THESE de DOCTORAT D'ETAT

ès Sciences Mathématiques

présentée

à l'Université Pierre et Marie Curie

- Paris VI -

par Monsieur François MARCOTORCHINO

pour obtenir le grade de DOCTEUR ès SCIENCES

Sujet de la thèse :

AGREGATION DES SIMILARITES EN CLASSIFICATION AUTOMATIQUE

soutenue le 25 Juin 1981

devant le jury composé de :

Messieurs	D. DUGUE	*(président)*
	P. DEHEUVELS	*(rapporteur)*
	J. GEFFROY	*(rapporteur)*
	G. KREWERAS	*(rapporteur)*
	M. BARBUT	*(examinateur)*
	I.C. LERMAN	*(invité)*

REMERCIEMENTS

Cette thèse présentée à l'Université Paris VI a été réalisée sous la direction du Professeur D. DUGUE, Directeur de l'I.S.U.P.. Que le Professeur D.DUGUE, avec qui j'ai poursuivi une grande partie de mes études en statistiques et qui avait il y a dix ans déjà accepté la présidence du jury de ma thèse de 3ème cycle, trouve ici une nouvelle fois l'expression de mes remerciements.

Je remercie très vivement le Professeur J.GEFFROY, Directeur du 3ème cycle "Statistiques et Recherche Opérationnelle" de l'Université Paris VI, d'avoir accepté la charge de rapporteur de la thèse et de m'avoir, à cette occasion donné la possibilité d'enseigner dans le cadre de son D.E.A..

Je suis d'autre part extrêmement reconnaissant au Professeur G. KREWERAS de l'Université Paris II, d'avoir accepté d'être, une fois encore, rapporteur d'une thèse que je présente. Ses critiques constructives m'ont, à nouveau, été fort utiles.

Je remercie le Professeur P. DEHEUVELS de l'Université Paris VI qui, ayant assisté à la progression de ces travaux et s'y étant favorablement interessé, a bien voulu accepter d'être également rapporteur.

Toute ma gratitude va aussi au Professeur M. BARBUT, Vice-Président de l'Université Paris V, l'un des premiers en France à s'être interessé à ce type de problème, pour sa participation au jury de cette thèse.

Mes remerciements vont également au Professeur I.C. LERMAN de l'Université de Rennes I d'avoir accepté l'invitation qui lui a été faite d'assister à la soutenance de cette thèse.

Par ailleurs, je tiens à profiter de l'occasion qui m'est donnée ici pour remercier à nouveau très chaleureusement Monsieur R. MOREAU, Directeur du Développement Scientifique IBM-France, d'avoir, depuis mon arrivée au sein de son Service, suivi avec le plus grand interêt le déroulement de mes travaux, de m'avoir incité à publier et judicieusement conseillé dans le choix d'applications de ces méthodes.

Je suis également très redevable à Monsieur M. PAPO, Directeur Scientifique de la Compagnie IBM-France, de la bienveillance qu'il a constamment manifestée et ce, dès l'origine, à l'égard de mes recherches en Analyse des Données et d'en avoir favorisé la réalisation tant sur le plan matériel que scientifique.

Je me dois d'associer à ces remerciements les autres personnalités scientifiques qui, les premières, se sont interessées à ces travaux. Je citerai ici plus particulièrement A.J. HOFFMAN du Laboratoire de Recherche IBM de Yorktown Heights, ainsi que B. MONJARDET de l'Ecole des Hautes Etudes en Sciences Sociales.

Pour terminer, mes remerciements vont à toutes les personnes qui, d'une façon ou d'une autre, ont permis la concrétisation de cette thèse; je pense ici en parlicilier aux spécialistes du système MPSX/370 M. BENICHOU, J.M. GAUTHIER, G. RIBIERE et bien entendu à mon collègue et ami P. MICHAUD.

UNIV.
LEIDEN
BIBL.

INTRODUCTION GENERALE

I - PRESENTATION DE LA THESE

Le problème dont il sera question dans cette thèse, résumé sous le vocable d'"Agrégation des similarités", n'est qu'un cas particulier de la modélisation générale de l'"agrégation Relationnelle" que j'ai proposée , conjointement avec mon collègue P. Michaud, dans le livre (R12).
En dépit de la ressemblance mathématique des modèles, l'importance du nombre des propriété spécifiques et les caractéristiques intrinsèques de l'Agrégation des Similarités justifient les développements présentés ici au sein d'une recherche plus personnelle menée parallèlement au travail de P. Michaud sur l'Agrégation des Préférences, présenté dans sa thèse (R13).

Cette introduction est essentiellement destinée à fournir des éclaircissements sur les points importants et originaux des textes (T1), (T2) et (T3) formant le corps de cette thèse.

La numérotation en (T.) renvoie aux textes suivants:
(T1): "Agrégation des Similarités en Classification Automatique";
(T2): "Optimisation en Analyse des Données Relationnelles";
(T3): "Optimization in Exploratory Data Analysis".
Une bibliographie dont les références sont notées (R.) figure à la fin de l'introduction, complétant ainsi les bibliographies propres à chacun des articles (T1), (T2), et (T3).

On trouvera, également adjointes à cette introduction, trois annexes dans lesquelles figurent certains résultats complémentaires dont la connaissance est implicitement supposée dans les différents textes de base.

Enfin, l'occasion me sera donnée dans cette introduction de préciser la part de travail personnel qui s'attache à l'obtention de certains des résultats originaux présentés ici.

II- SITUATION DU PROBLEME DE L'AGREGATION DES SIMILARITES

L'agrégation des similarités , présentée dans cette thèse est une généralisation du problème de la recherche d'une "partition centrale", formalisé par S. Regnier en 1965 dans (R15).
Le problème de la recherche d'une partition "centrale" peut se présenter de la façon suivante:

Un ensemble de m "juges" (qui pourront être également des "critères") donnent leurs opinions sur une population de n objets en termes de partitions de cette population; rechercher alors la partition "centrale" consiste à trouver la partition des objets la plus "proche" au sens de la distance de la "différence symétrique" de chacune des partitions individuellement donnée par les m juges. (Page 9 à 11 du texte T1 on montre que ceci est équivalent à utiliser une règle majoritaire d'agrégation, en l'occurence la règle de Condorcet).

A titre d'illustration considérons cet exemple type de recherche de "partition centrale".
On suppose ici que 5 personnes ont été interrogées sur la façon dont elle percevaient les parentés politiques de 6 des candidats de l'élection présidentielle de Mai 81. Leurs réponses sont répertoriées dans le tableau no1:

	J1	J2	J3	J4	J5
VGE	1	1	1	1	1
CHIR	1	2	1	2	1
MITT	2	3	2	3	2
MARC	2	4	3	3	2
DEBR	1	2	4	2	1
CREP	2	3	2	3	1

Tableau n1

	VGE	CHIR	MITT	MARC	DEBR	CREP
VGE		3	0	0	2	1
CHIR	3		0	0	4	1
MITT	0	0		3	0	4
MARC	0	0	3		0	2
DEBR	2	4	0	0		1
CREP	1	1	4	2	1	

Tableau n2

Dans le tableau n1, on voit que lorsque deux candidats ont été jugés "similaires" par un juge ils ont la même valeur du code de classification.(Exemple pour le Juge 1 VGE ,CHIRac et DEBRé ont été jugés "similaires", ils ont tous les trois le code "1"; c'est à dire ils sont dans une même classe).
Dans le tableau n2 (qu'on appelera par la suite tableau de comparaisons par paires), on a compté pour chaque couple de candidats (i,j), le nombre de "juges" qui les ont trouvés similaires. Exemple: Trois "juges" ont trouvé que VGE et CHIRac étaient "semblables" d'un point de vue image politique.
Parmi les partitions "centrales", solutions du problème précédent (il peut en effet y avoir plusieurs solutions) citons: P1= (VGVE.,CHIR.,DEBR.)-(MITT.,MARCH.,CREP.)
P2= (V.G.E.)-(CHIR.,DEBR.)-(MITT.,CREP.)-(MARCH.)
Pour ces deux solutions la valeur des fonctions critères que nous présenterons plus loin est la même. (Les moyens méthodologiques permettant d'obtenir lesdites partitions seront présentés ultérieurement).

Ce petit exemple n'avait pour but que d'illustrer la façon de poser un problème de partition centrale sous sa forme la plus simple.
Le concept d'"agrégation de similarités" généralise ce problème de "partition centrale" à des données initiales ne représentant pas des partitions.

Le problème particulier de la recherche d'une partition centrale a ,quant à lui, été étudié essentiellement par des chercheurs français. Outre S. Regnier (R15 et R16), on trouve le problème présenté à la fois dans le tome I de "l'Analyse des Données" de J.P. Benzecri et al. (R2), dans le livre de I.C. Lerman (R9), dans les travaux de De la Vega (R17). Seul auteur étranger à s'immiscer dans ce concert français, B. Mirkin (R14) a posé le problème de la Partition Centrale approximativement au même moment que S. Regnier, mais de façon moins détaillée.
Cela dit, avant Regnier un certain nombre d'auteurs - je pense ici à M. Barbut (R1) et C. Zahn (R20)- avaient vu le rôle de la distance de la "différence symétrique" dans les problèmes de proximité entre relations binaires.

Dans un tel contexte, l'apport essentiel de cette thèse a été de pouvoir résoudre de façon exacte le problème de la mise en évidence de la (ou des) partition(s) centrale(s); problème qui était souvent considéré par la plupart des auteurs comme impossible à résoudre pratiquement.
Pour s'en convaincre, il suffit de consulter S. Regnier (R16 p.23), J.P. Benzecri (R2 p.237), I.C. Lerman (R9 p.XI), M. Jambu (R8 p.239), l'introduction de la thèse de M. Roux (R18), etc ...

La méthodologie présentée ici généralisera au cas de données non nécessairement issues de partitions la recherche d'une partition solution qui sera également "Centrale" d'un point de vue métrique.
Au niveau conceptuel, l'agrégation des similarités joue un rôle relativement excentré par rapport aux multiples méthodes proposées en Classification Automatique.
En effet, on peut diviser les méthodes de Classification Automatique en deux grands groupes: les méthodes hiérarchiques d'une part, les méthodes non-hiérarchiques d'autre part.
Les méthodes hiérarchiques, dont les meilleurs algorithmes sont présentés par M. Jambu (R8), ont essentiellement comme qualité de permettre la visualisation, sous forme d'arbre de classification, des proximités relatives entre les objets. Cet avantage au niveau de la "visualisation" peut également être considéré comme un défaut car il y a autant d'arbres de classification pour un problème donné que de méthodes hiérarchiques potentiellement utilisables sur ces mêmes données (nécéssitant soit a priori, soit a posteriori une certaine subjectivité).
A l'opposé se trouvent les méthodes non-hiérarchiques d'obtention d'une partition finale, dont les tenants sont en France E. Diday (R5) avec la fameuse et très utilisée "Méthode des Nuées Dynamiques", I.C. Lerman et H. Lerrede (R11) avec leur "Méthode des pôles". Ces méthodes, bien que très rapides et moins encombrantes en taille mémoire que les

précédentes, nécessitent une hypothèse d'applicabilité préalable, à savoir la fixation a priori du nombre de classes de la partition cherchée.

La méthode proposée ici ne présente pas le défaut évoqué plus haut, tout en proposant un type de résultat identique, ce qui de ce fait la rattache à la catégorie "non-hiérarchique".
Ceci étant posé la méthode d'"Agrégation des Similarités", même dans sa version heuristique, n'a pas les possibilités de rapidité qu'offrent les "nuées dynamiques" ou les méthodes d'agglomération hiérarchique, limitant ainsi d'un point de vue pratique le nombre d'objets à classifier à n = 200.

III ANALYSE DES ARTICLES (T1), (T2), (T3)

A)- ANALYSE DU TEXTE (T1)

Le texte (T1) étant en lui-même une synthèse, certains résultats donnés ici et non démontrés renvoient à des articles antérieurs. Cependant, ce paragraphe de commentaire me permettra à la fois d'expliciter certains résultats originaux figurant dans (T1) et de donner des précisions sur l'environnement méthodologique existant.

L'apport original du texte (T1), relativement aux autres articles écrits antérieurement par P. Michaud et moi-même, peut se caractériser par:
- la présentation synthétique des modèles d'Agrégation des Préférences et des Similarités (et plus particulièrement de la relation avec la règle de Condorcet);
- la démonstration d'une condition d'admissibilité de la règle de Condorcet en Agrégation des Similarités (condition d'emboitement);
- la présentation des propriétés structurelles vérifiées par la méthode d'Agrégation des Similarités, soit:
    - propriété de Consistance Additive
    - propriété de Neutralité Totale
    - propriété de de non-fixation du nombre de classes de la partition solution Y;
- la mise en évidence de la non-vérification de la propriété de consistance additive, pourtant très intuitive, par certaines méthodes de classification hiérarchique;
- l'attention apportée à la "prise en compte des données", qu'il s'agisse de données de partition ou de données de présence-absence (relation entre les indices usuels et la règle de Condorcet, codage des données du type Czekanowski etc..);
- la définition d'indices généralisés et modifiés de concordance;

- la présentation d'indices de cohérence et interprétation statistique;
- et pour terminer, 3 nouveaux exemples d'Agrégation des Similarités seront présentés.

a)- Remarques sur les tableaux de Comparaisons par Paires (T1 p.6)

Le tableau de Comparaisons par paires C^k qui sert au codage des données individuelles est assimilé à une relation binaire dans l'ensemble des articles que nous avons écrits avec P. Michaud. Cette assimilation s'effectue au détriment de la réflexivité comme à celui de la possibilité de boucle sur chaque objet. D'un point de vue mathématique le tableau C^k ne peut représenter "toutes les relations binaires" puisque sa diagonale n'est pas considérée, mais pour les calculs que nous devons effectuer pratiquement cette restriction n'a strictement aucune conséquence, si ce n'est de simplifier le problème.
D'autre part, si à l'origine C^k est considéré comme une "relation" en posant $c^k_{ij} = 1$ ou $c^k_{ij} = 0$, on a généralisé le codage de C^k aux configurations de "non-réponses" en proposant le codage 0, 1/2, 1 donné page 27; ce qui nous oblige donc de ce fait à ne plus considerer C^k comme une "relation binaire", mais offre par ailleurs des possibilités plus importantes dans la prise en compte des données.

b)- Remarque sur le critère d'agrégation (T1 p.9)

Comme il est dit en (T1 p.9), le critère d'agrégation est la formulation linéaire de la règle de Condorcet:

$$F(Y) = \sum_{ij \neq i} \sum c'_{ij} \; Y_{ij}$$

Le choix de ce critère est du aux propriétés qu'il vérifie, présentées ultérieurement dans (T1 pp. 21, 22, 23). Néanmoins, toute fonction séparable des c'_{ij} et de Y pourrait servir de critère. On peut donc proposer dans ce type d'approche la fonction critère définie de la façon suivante:

$$F'(Y) = \sum_{i} \sum_{j \neq i} Sgn\,(c'_{ij}) \; Y_{ij}$$

où $Sgn\,(c'_{ij}) = 1$ si $c'_{ij} > 0$,
$= 0$ si $c'_{ij} = 0$,
$= -1$ si $c'_{ij} < 0$.

Cette fonction, liée par une procédure d'écrêtage des majorités de Condorcet par paires, est dérivée de la procédure de "Copeland" définie pour les problèmes d'agrégation des préférences (procédure où l'on ne considère que les victoires, défaites ou match-nuls dans les confrontations par paires).

c)- Remarque sur "l'Effet Poincaré" (T1 p.11)

Nous avons appelé dans (T1) "Effet Poincaré" le principe d'intransitivité de la perception collective de la "Similarité" par analogie avec la notion d'"Effet Condorcet". Cette attribution à Poincaré d'un paradoxe sur lequel il n'a pas travaillé pourrait sembler abusive s'il n'avait, dans ses écrits sur le "continu physique" dans Science et Hypothèse en 1907 (approfondi en 1908 par E. Borel voir (R3)), introduit la notion de "non discernabilité" assimilable à la notion d'indifférence dans les quasi-ordres; notion elle même très analogue à la notion de similarité. Cela étant, le "sophisme du tas de blé" des anciens grecs est lié à cette perception de la non transitivité de la ressemblance.

d)- Remarque sur la taille des problèmes (T1 p.12)

Puisque le problème d'agrégation des similarités consiste à trouver une partition, il est évident qu'a priori Y peut être l'une des B(n) partitions des n objets à classifier.
B(n) étant le nombre de Bell = $\sum_k S_n^k$

où S_n^k = Nombre de partitions de n objets en k classes
= Nombre de Stirling de deuxième espèce;

la structure combinatoire de B(n) est évidente car:

$$B(n) = \frac{1}{e} \sum_{k=1}^{\infty} \frac{k^n}{k!}$$

Un calcul utilisant la récurrence sur les S_n^k, nous permet d'obtenir en APL les valeurs suivantes jusqu'aux limites du dépassement de capacité, c'est à dire pour n = 71:

$$B(71) = 4{,}08113.10^{74} \quad ,$$

nombre énorme qui montre bien qu'une exploration exhaustive des B(n) partitions est impossible dès que n>25, B(25) étant égal à $1{,}203.10^{18}$; et ce, même avec les ordinateurs les plus puissants.
Il n'en reste pas moins vrai que l'approche qui consiste à fixer le nombre de classes de la partition à trouver (procédé qui se justifie pleinement quand on a une certaine connaissance de la population à classifier), conserve au problème un caractère hautement combinatoire et ne le simplifie pas.

En effet on a par exemple:

$$\max_k S_{71}^k = 7{,}57079.10^{73}$$

(nombre très peu différent, du point de vue importance, du

nombre total partitions)

on peut montrer que

$$\frac{\underset{k}{\text{Max}}\ S_n^k}{B(n)}$$

décroît irrégulièrement et en tendance seulement avec n, de n=2 jusqu'à n=71. On a:

$$\underset{k}{\text{Max}}\ S_n^k > 0{,}184\ B(n)$$

D'autre part on a $S_n^k > 10^{40}$ pour $4 \leq k \leq 54$, $n = 71$

donc le problème est de nature combinatoire pour toute fixation du nombre de classes comprise entre 4 et 54 (pour n=71 par exemple).

e)-Remarque sur la CONDITION d'EMBOITEMENT (T1 p.13)

J'ai donné page 13 de T1 une condition que doivent vérifier les tableaux de similarités individuels C^k pour que l'application de la Règle de Condorcet en Agrégation des Similarités donne une solution admissible, cette condition est appelée "Condition d'emboîtement". Cela dit, il existe d'autres configurations où la Règle de Condorcet peut s'appliquer sans donner lieu à des "Effets Poincaré".En particulier des configurations triviales pour lesquelles on dispose d'un nombre impair de juges donnant leurs similarités sous forme de deux modèles de partitions possibles seulement (la solution de l'agrégation étant alors l'une de ces deux partitions possibles).

f)- Remarque sur les temps de calcul de la méthode heuristique (T1 p.21)

Nous donnons page 21 un tableau des temps de calcul sur un échantillon de 18 exemples réels traités d'une part par la méthode exacte, d'autre part par la méthode heuristique. Contrairement au cas de l'agrégation des préférences où l'heuristique figure dans le livre (R12), l'heuristique en agrégation des similarités n'est présentée dans aucun des textes cités. C'est pourquoi on en trouvera dans (ANNEXE 1) une présentation succincte, mais suffisamment explicite pour être programmée.

g)- Remarque sur les propriétés structurelles de la méthode (T1 p.23)

Outre les trois propriétés présentées en (T1 p.23), la méthode d'agrégation des similarités vérifie la propriété d'unanimité Paretienne que j'ai rajoutée en (ANNEXE 2).

Il importe d'insister sur la propriété numéro 3 de "NON FIXATION DU NOMBRE DE CLASSES" de la partition cherchée. Cette propriété est la caractéristique majeure de cette méthodologie. Les méthodes usuelles d'obtention de partition qu'elles soient françaises, E. Diday, ou anglo-saxonnes, Freeman, Forgy (R7 et R8) etc..., nécessitent la donnée préalable du nombre de classes de la partition solution.
De ce fait, les griefs méthodologiques faits par bon nombre de spécialistes de la taxinomie hiérarchique aux tenants de la classification non hiérarchique (voir M. Jambu (R8 p.242)) à propos de l'obligation de fixer le nombre de classes des partitions, ne peuvent être invoqués ici.

h)- Remarque sur la prise en compte des données (T1 pp. 25 à 44)

Bien que le chapitre sur la prise en compte des données soit le plus important et le plus développé du texte (T1), certaines précisions sont nécessaires pour distinguer la méthode d'agrégation des similarités des méthodes de classification existantes.

En effet la méthode d'agrégation des similarités s'applique, avec une efficacité maximum et dans toute l'acception du terme, aux données (c'est à dire à des tableaux de comparaisons C^K) représentant directement des partitions ou des partitions empiétantes. Ces données correspondent alors à des variables à modalités; conditions qui, comme nous le savons, nécéssitent un effort de recodage lors de l'application des autres méthodes ((T1) p.30).

Réciproquement, lorsque les données sont issues de mesures, de comptages ou de prise en compte de variables quantitatives ((T1) p.32), il est nécessaire de passer par l'intermédiaire d'histogrammes descriptifs de chaque variable pour pouvoir appliquer la méthode d'Agrégation des Similarités, contrairement à certaines méthodes de classification hiérarchiques par agglomération successive des classes: Méthode de Ward (R19) avec métrique du χ^2, ou méthode Centroide avec métrique euclidienne par exemple.
Pour ces dernières, en effet, la notion de similarité entre les objets n'a pas besoin de se définir "intrinsèquement", mais "relativement" par rapport aux écarts "en distance" entre les objets représentés sur un arbre de classification; et c'est la vision "globale" de l'arbre de classification qui permet alors de se faire une idée de la parenté relative des objets les uns par rapport aux autres.

Restent les données binaires. Dans ce cas, on constate qu'en fonction de l'interprétation donnée aux configurations (1-1) et (0-0), le codage par tableaux de comparaisons en Agrégation des Similarités permet de reconsidérer la plupart des indices de similarité usuels comme de simples variantes des règles de majorité par paires (majorité simple, au tiers, au deux-tiers, etc...(Marcotorchino)).
D'ailleurs, à cette occasion j'ai proposé un nouvel indice possédant un certain nombre de propriétés parmi lesquelles on peut tout particulièrement citer la neutralité des configurations (0-0) vis à vis de la règle de Condorcet.

Ce chapitre concerne également le codage de certaines données spéciales, en particulier les tableaux de Bertin-Czekanowski dont je donne l'interprétation en termes de juges (T1 p. 28).

i)- Remarque sur l'indicateur généralisé de concordance (T1 p.44)

Cet indice de concordance introduit pour mesurer la qualité des données d'agrégation est valable quel que soit le problème d'agrégation.
Il varie (sous la forme donnée p.44)
de 0 à 1 pour m pair,
de $1/m^2$ à 1 pour m impair.
Un certain nombre de ses caractéristiques sont plus particulièrement étudiées en (ANNEXE 3).

j)- Remarque sur l'indice généralisé de cohérence (T1 p.46)

P. Michaud donne dans sa thèse (R13), outre sa forme générale, le mode d'obtention des Max F(Y) pour $Y \in RP$ permettant une formulation directe de cet indice.
On peut également fournir une expression explicite du dénominateur

$$D(Y) = \max_{Y \in RP} F(Y) - \min_{Y \in RP} F(Y)$$

pour les 5 types de relations RP{To - AS - To∩AS - S - Q} et déduire alors une interprétation statistique de cet indice dans les cas où C' est antisymétrique ou symétrique. (Marcotorchino)
On a en effet:

1) $D(Yo) = 1/2 \sum_{i<j} [|c'_{ij} - c'_{ji}| + |c'_{ij}| + |c'_{ji}|]$ pour $Yo \in To$

2) $D(Yo) = 1/2 \sum_{i<j} [|c'_{ij} - c'_{ji}| + |c'_{ij}| + |c'_{ji}|]$ pour $Yo \in AS$

3) $D(Yo) = \sum_{i<j} |c'_{ij} - c'_{ji}|$ pour $Yo \in To \cap AS$

4) $D(Yo) = \sum_{i<j} |c'_{ij} + c'_{ji}|$ pour $Yo \in S$

5) $D(Yo) = \sum_{i<j} Max\ (|c'_{ij} + c'_{ji}|\ ,\ |c'_{ij} - c'_{ji}|)$ pour $Yo \in Q$

On constate ici que l'intervalle de variation
$D(Yo) = Max\ F(Yo) - Min\ F(Yo)$
a la même valeur pour $Yo \in To$ et $Yo \in AS$

Dans le cas où les tableaux C' sont antisymétriques ou symétriques on a la simplification suivante pour D(Yo) dans les cas $AS \cap To$ et S:

$D(Yo) = 2 \sum_{i<j} |c'_{ij}|$ pour $Yo \in AS \cap To$ et C' antisymétrique

$D(Yo) = 2 \sum_{i<j} |c'_{ij}|$ pour $Yo \in S$ et C' symétrique.

Le numérateur de l'indice d'incohérence N(Yo,Y) vaut alors:

$N(Yo,Y) = \sum_{i<j} |c'_{ij}| - \sum_{i \neq j} c'_{ij} y_{ij}$ pour $Yo \in AS \cap To$ et $Y \in AS \cap To \cap T$

$N(Yo,Y) = \sum_{i<j} c'_{ij} + \sum_{i<j} |c'_{ij}| - \sum_{i \neq j} c'_{ij} y_{ij}$ pour $Yo \in S$ et $Y \in S \cap T$

Avec les notations statistiques de (T1) page 48 ces quantités deviennent:

$D(Yo) = 2\ (N1+P1)$ pour $Yo \in AS \cap To$ et $Y \in AS \cap To \cap T$ et C' antisymétrique.

$D(Yo) = 2(N1+P1+P0+N0)$ pour $Yo \in S$ et $Y \in S \cap T$ et C' symétrique

De la même façon on a :

$N(Yo,Y) = N1+P1-(P1-N1) = 2N1$ pour $Yo \in AS \cap To$ et $Y \in AS \cap To \cap T$, C'antisym.

$N(Yo,Y) = N1+P1+P0+N0+P1+P0-(N1+N0)-2(P1-N1)$
$= 2(N1+P0)$ pour $Yo \in S$ et $Y \in S \cap T$, C' sym.

On constate donc que dans les deux cas l'indice d'incohérence s'exprime simplement en fonction de N1,N0,P1,P0 suivant les formules:

$I(incoh) = N1/(N1+P1)$ pour $Yo \in AS \cap To$, $Y \in AS \cap To \cap T$ et C' antisymétrique.

$I(incoh) = (N1+P0)/(N1+P0+P1+N0)$ pour $Yo \in S$ et $Y \in S \cap T$ et C' symétrique.

B) - ANALYSE DE (T2)

L'apport de l'article (T2) par rapport au texte (T1) réside d'une part essentiellement dans la présentation relativement complexe du programme dual associé au modèle linéaire de l'Agrégation des Similarités; d'autre part dans la présentation d'un exemple portant sur des données de similarités directes montrant les possibilités d'élargissement de "l'Agrégation de Similarités Générale" par rapport à la seule "recherche d'une Partition Centrale".

a)- En effet, page 12 de l'article (T2) on trouve l'expression du dual pour le problème d'Agrégation des Similarités, dual calculé à partir de la forme inégalité du problème (donnée page 11).

La caractéristique de la forme duale calculée à partir du primal mis sous forme inégalité est de n'avoir que $n(n-1)/2$ contraintes au lieu des $0(n^3)$ propres à la forme standard, taille qui bloque à priori toute tentative de calcul pratique par l'algorithme du simplexe en raison de l'encombrement mémoire qui en résulterait.

On trouvera dans (T1) pages 19 à 21 un exposé plus détaillé sur le principe d'application de la méthode exacte par le code IBM MPSX/370.

b)- L'exemple présenté pages 14 et 15 est relatif aux données de similarités directes. Il faut entendre par là des données issues de comparaisons par paires directes et non obtenues à partir de données de partitions ou de pseudo-partitions.

sous cette forme (voir (T1) pages 26 et 27), les données sont directement assimilables par l'algorithme "d'Agrégation des Similarités".

Dans ce cas chaque juge Jk (il y en a 57 dans l'exemple cité) donne ses similarités propres sur les 11 objets à classifier sous la forme d'un tableau de similarités C^k individuel. Chaque tableau binaire C^k peut très bien ne pas être transitif, donc ne pas correspondre à une partition. Néanmoins, le tableau collectif

$$C = \sum_{k=1}^{57} C^k$$

donné page 15 est directement transformable en un tableau

$$C' = 2C - 57$$

et ce sont les valeurs

$$c'_{ij} = 2\, c_{ij} - 57$$

qui rentreront dans l'expression de la fonction économique à maximiser:

$$F(y) = \sum_{i \neq j} c'_{ij} y_{ij}$$

Les résultats de la partition optimale sont donnés page 15. Cet exemple souligne ici le caractère plus général de "l'Agrégation des Similarités" par rapport au problème de "Partitions Centrales", puisque nous agrégeons ici des données qui ne correspondent pas à des partitions.

C)- ANALYSE DE (T3)

Ce troisième article a été présenté au IIIème Congrès International de Recherche Opérationnelle (KOLN, 1980). Reprenant en grande partie les points actifs de la méthode (décrits dans (T1)) il présente cependant deux particularités:
- il détaille un exemple de "sériation" en archéologie;
- il énumère un certain nombre de propriétés propres aux modèles d'agrégation relationnelle.

- L'étude des problèmes de "sériation"

Par "sériation" (en archéologie) on entend la mise en évidence, par des moyens mathématiques adaptés au traitement d'un tableau de données, d'un phénomène sous-jacent en évolution (par exemple une chronologie) et la variable qui conditionne cette évolution (par exemple le temps).
Rappelons avec I.C. Lerman ((R10), page 39) que l'hypothèse valable dans les problèmes de sériation peut être formulée de la façon suivante:
"Un type d'objet donné existe pendant une période continue de temps; de plus si at1, at2, at3 sont trois types apparus aux dates t1, t2 et t3 avec t1 < t2< t3 le type at2 est plus "proche" de at1 que ne l'est at3, respectivement le type at2 est plus "proche" de at3 que ne l'est at1".

Dans le cas présent on se servira de l'hypothèse duale, en définissant la similarité entre deux sites archéologiques par le nombre d'objets ou d'outils qui s'y retrouvent avec des fréquences voisines.
Le but recherché ici est la mise en évidence de la "sériation" sur les sites par la seule agrégation des similarités ainsi définies. La partition optimale est d'ailleurs en accord avec les résultats obtenus par l'analyse des correspondances sur le tableau de "contingences" des données brutes.

IV - BIBLIOGRAPHIE DE L'INTRODUCTION GENERALE

(R1) M.BARBUT,"Médiane, Distributivité, Eloignements", Publications du Centre de Mathématiques Sociales, E.P.H.E.,6ème section, (1961).

(R2) J.P.BENZECRI et coll, L'Analyse des Données, Tome I: la Taxinomie, Tome II: l'Analyse des Correspondances, Paris, Dunod, (1973).

(R3) E. BOREL, "Probabilité et Certitude", Paris, P.U.F. (1950).

(R4) E.DIDAY, "Nouvelle méthode et nouveaux concepts en Classification Automatique et Reconnaissance des Formes", Thèse de Doctorat d'Etat, Université Paris VI, (1972).

(R5) E. DIDAY, "La méthode des nuées dynamiques", Revue de Statistiques Appliquées, Vol. 19, no.2,(1971), pp.19-34.

(R6) E.W. FORGY, "Cluster Analysis of multivariate data", efficiency versus interpretability of classifications (abstracts), Biometrics 21, (1965), p.768.

(R7) M. FREEMAN, "Experiments in discrimination and classification", Pattern recognition, Vol.1, no.3,(1969).

(R8) M.JAMBU, "Classification automatique pour l'analyse des données". Tome I: méthodes et algorithmes, avec M.O.LEBEAUX. Tome II: Logiciels, Dunod (1978).

(R9) I.C.LERMAN, "Les Bases de La Classification Automatique", Gauthier Villars, Paris, (1970).

(R10) I.C.LERMAN. "Analyse du Phénomène de la Sériation à partir d'un tableau d'incidence", Mathématiques et Sciences Humaines, n0 38, (1972), p.39.

(R11) I.C.LERMAN, "Reconnaissance et classification des structures finies en analyse des données", Vol. I, Théorie et Méthodes, Rapport no 70, IRISA, Université de Rennes, (1977).

(R12) F.MARCOTORCHINO et P.MICHAUD, "Optimisation en Analyse Ordinale des Données", Masson, Paris, (1979).

(R13) P.MICHAUD, "Agrégation des Préférences", Thèse d'Etat présentée à l'Université Paris VI, (1981).

(R14) B.MIRKIN, "Group Choice", édité par P. Fishburn, V.Winston and Sons, J.Wiley, (1979).

(R15) S.REGNIER, "Sur Quelques Aspects Mathématiques des Problèmes de Classification Automatique", I.C.C. Bulletin, Rome, (1965).

(R16) S. REGNIER, "Stabilité d'un opérateur de Classification", Mathématiques et Sciences Humaines, no 60, (1977), p.23-30.

(R17) S. REGNIER, W. de la VEGA, "Préhension et interprétation de plusieurs classifications d'un même ensemble de données", compte-rendu de contrat D.G.R.S.T., Paris, A.D.I.S.H., (1976).

(R18) M. ROUX, "Un algorithme pour construire une hiérarchie particulière", Thèse de 3ème Cycle, Université Paris VI, (1968).

(R19) J.H. WARD, "Hierarchical Grouping to Optimize an Objective Function", Journal of American Statistical Association, no 58, (1963), pp. 236-244.

(R20) C.T.ZAHN, "Approximating Symmetric Relations by Equivalence Relations", SIAM Journal of Applied Mathematics, no 12, p. 840, New York, (1964).

ANNEXE 1

(PRESENTATION DE L'HEURISTIQUE)

Ce texte est une présentation succincte de l'heuristique que nous avons utilisée parallèlement à la méthode exacte dans le cas de "l'agrégation des similarités", et dont certains temps de calculs sont donnés dans (TI) page 21.

Cette heuristique repose sur l'utilisation séquentielle et récursive de 4 conditions d'amélioration possible de la fonction économique à maximiser:

$$F(Y) = \sum_{ij \neq i} \sum c'_{ij} y_{ij} \quad (\text{ici } c'_{ii} = 0 \quad \forall\, i\,);$$

ces conditions étant différentes et indépendantes les unes des autres.

CONDITION 1: Condition des "transferts"

La première Condition est la condition dite des "transferts" ou des "échanges"; elle est due à S.Regnier (voir (R15)).

Supposons que nous partions d'une partition de n objets en p classes $\mathcal{C}_k$.
Alors on appelle :

$$C_i(k) = 2 \left(\sum_{j \in \mathcal{C}_k} c'_{ij} \right) = \text{Contribution d'un objet i à la classe } \mathcal{C}_k$$

La condition de "non transfert" stipule que la contribution d'un objet i à sa propre classe doit être supérieure à sa contribution à toutes les autres classes. Ce qui se traduit par:

$$C_i(i) \geq C_i(\ell) \quad \forall\, \mathcal{C}_\ell \neq \mathcal{C}_i$$

C'est à dire $\sum_{j \in \mathcal{C}_i} c'_{ij} = \operatorname{Max}_{\ell} \sum_{j \in \mathcal{C}_\ell} c'_{ij}$

En d'autres termes, il y aura transfert possible de l'objet i de sa classe $\mathcal{C}_i$ à la classe $\mathcal{C}_\ell$ si :

$$C_i(\ell) - C_i(i) > 0$$

le gain dans la fonction économique sera alors égal à:

$$\Delta(i) = C_i(\ell) - C_i(i) = 2 \left(\sum_{j \in \mathcal{C}_\ell} c'_{ij} - \sum_{j \in \mathcal{C}_i} c'_{ij} \right)$$

Après un transfert on est dans la nouvelle situation (NC étant le nombre de classes de la partition avant transfert)

$$NC := NC$$

$$\mathcal{C}_i := \mathcal{C}_i - \{i\} \quad \text{et} \quad \mathcal{C}_\ell := \mathcal{C}_\ell + \{i\}$$

CONDITION 2: Condition de "positivité"

Comme extension de cette condition des "Transferts" on peut ajouter la condition dite de positivité:
la contribution d'un point à sa classe doit être positive ou nulle.

soit $C_i(i) \geq 0 \Rightarrow \sum_{j \in \mathcal{C}_i} c'_{ij} \geq 0$

Dans le cas contraire si:
$C_i(\ell) < 0 \quad \forall \ \ell$
alors on isolera "i" puisque par convention $c'_{ii} = 0$
d'où $C_i(i) \geq C_i(\ell)$ car $C_i(i) = c'_{ii}$
et $\{i\}$ sera une classe de la partition à lui tout seul.
Dans ce cas, si on isole un point i on obtient alors:
$NC := NC+1$ et $\mathcal{C}_\ell := \mathcal{C}_\ell - \{i\}$

Réciproquement s'il existe un indice ℓ pour lequel $C_i(\ell) > 0$ on transfère le point $\{i\}$ dans la classe d'indice ℓ. De ce fait la nouvelle situation est donnée par:

$$NC := NC - 1 \quad \text{et} \quad \mathcal{C}_\ell := \mathcal{C}_\ell + \{i\}$$

CONDITION 3: Condition "d'agglomération"

La troisième condition est la condition d'agglomération de deux classes.
En effet, supposons qu'à une étape donnée de l'heuristique on ait deux classes $\mathcal{C}_k$ et $\mathcal{C}_\ell$ telles que:

$$\sum_{i \in \mathcal{C}_k} \sum_{j \in \mathcal{C}_\ell} c'_{ij} > 0$$

alors par application de cette condition, on réunira les classes $\mathcal{C}_k$ et $\mathcal{C}_\ell$ en une même et unique classe ($\mathcal{C}_k \cup \mathcal{C}_\ell$) ; ainsi le gain dans la valeur de la fonction économique sera égal à:

$$\Delta(\mathcal{C}_k, \mathcal{C}_\ell) = 2 \left(\sum_{i \in \mathcal{C}_k} \sum_{j \in \mathcal{C}_\ell} c'_{ij} \right)$$

Dans ce cas on a $NC := NC - 1$
A ces conditions, j'ai ajouté les conditions supplémentaires figurant sous la rubrique Condition 4.

CONDITION 4: Condition "d'échanges simultanés" (Marcotorchino)

Elle peut être intitulée condition des "doubles transferts" ou "d'échanges simultanés".
Cette condition, jointe aux conditions précédemment utilisées, est comme on le verra sur de petits contre-exemples une condition de "déblocage" possible de la condition de "non transfert".
Elle consiste à transferer un objet "i" de sa classe $\mathcal{C}_i$ vers une classe $\mathcal{C}_j$ simultanément au transfert d'un objet "j" de sa classe $\mathcal{C}_j$ vers la classe $\mathcal{C}_i$, alors qu'aucun transfert n'est possible séparément.
La condition de "non transfert" de i de $\mathcal{C}_i$ vers $\mathcal{C}_j$ s'exprime par:

$$C_i(i) - C_i(j) > 0 \qquad (a)$$

La condition de "non transfert" de j de $\mathcal{C}_j$ vers $\mathcal{C}_i$ s'exprime par:

$$C_j(j) - C_j(i) > 0 \qquad (b)$$

Si par contre on pratique simultanément un "double transfert" la variation de la fonction économique est donnée par:

$$\Delta(i,j) = C_i(j) - C_i(i) + C_j(i) - C_j(j) - 4c'_{ij}$$

variation due aux transferts plus élimination du "pivot" c'_{ij}

soit:

$$\Delta(i,j) = 2\left(\sum_{\ell\in\mathcal{C}_j} c'_{i\ell} - \sum_{\ell\in\mathcal{C}_i} c'_{i\ell} + \sum_{\ell\in\mathcal{C}_i} c'_{j\ell} - \sum_{\ell\in\mathcal{C}_j} c'_{j\ell}\right) - 4c'_{ij}$$

Donc si le "pivot" c'_{ij} est suffisamment négatif on peut avoir $\Delta(i,j) > 0$, même si séparément les conditions de non transferts indépendants (a) et (b) sont vérifiées, il suffit pour cela que:

$$-4c'_{ij} > C_i(i) - C_i(j) + C_j(j) - C_j(i)$$

En d'autres termes, on peut fort bien effectuer des "doubles transferts" même si séparément des transferts simples sont impossibles.

Dans le cas présent, on a NC := NC
mais $\mathcal{C}_i := \mathcal{C}_i - \{i\} + \{j\}$
et $\mathcal{C}_j := \mathcal{C}_j - \{j\} + \{i\}$

On a testé par ailleurs des conditions plus complexes que la condition des "doubles transferts simultanés", comme la condition des "quadruples transferts" consistant à échanger 2 points d'une classe avec 2 points d'une autre, etc...
On obtient des possibilités d'amélioration de F(Y), mais à partir de là le coût de calcul devient prohibitif du fait de l'aspect combinatoire de l'inspection des 4-uples à pivots négatifs.

PRESENTATION DE L'ALGORITHME HEURISTIQUE

Cet algorithme se rattachera aux trois opérations indépendantes issues des conditions 1,2,3 et 4 définies précédemment.
Posons: [T] = opération de transfert (1) et d'isolement (2),
[A] = opération d'agglomération (3),
[ES] = opération d'échanges simultanés (4).
Munis de ces trois opérations séparées, on peut définir différents algorithmes récursifs fondés sur les conditions d'amélioration de F(Y); un principe identique a d'ailleurs été utilisé pour l'heuristique en Agrégation des Préférences, pour laquelle j'ai proposé également des "conditions locales d'optimalité" (voir (R12) pp.150-151).

L'algorithme peut être construit selon le schéma suivant:
①- Appliquer [T] jusqu'à vérification de la condition de non transfert,
②- Appliquer [A] jusqu'à vérification la condition de non agglomération,
③- Appliquer [ES] jusqu'à vérification de la condition de non échanges simultanés,
④- Aller en 1 si une amélioration pour F(Y) est intervenue dans la séquence ① - ② - ③, sinon aller en ⑤
⑤- Fin.

Ce même algorithme peut également se noter:
1 [T] --→ 2 [A] --→ 3 [ES] --→ 1 [T] --→ etc..

Toute autre permutation de [T] [A] et [ES] est possible comme:

1 [A] --→ 2 [T] --→ 3 [ES] --→ 1 [A] --→ etc..

1 [T] --→ 2 [ES] --→ 3 [A] --→ 1 [T] --→ etc..

En comparant les résultats de ces différents algorithmes heuristiques (sur plus de 180 exemples) avec les solutions exactes que nous avions obtenues indépendemment par la méthode exacte de programmation linéaire (voir (TI pages 19,20, 21) et ((R12) pp140 144)), nous sommes arrivés à la conclusion que les meilleurs algorithmes heuristiques sont du type:

I : 1 [A] --→ 2 [T] --→ 3 [ES] --→ 1 [A] --→ etc...
ou du type:
II : 1 [A] --→ 2 [T] --→ 3 [ES] --→ 2 [T] --→ 1 [A] --→ etc...
Les résultats obtenus par ces heuristiques avoisinaient 95% de solutions exactes dans les deux cas; résultat meilleur statistiquement que pour les autres combinaisons des opérations [A], [T], [ES].

Enfin, on peut signaler que chaque opération seule : [T], [A] ou [ES] donnait des résultats nettement moins bons.

En conclusion, l'heuristique que nous avons choisie est l'heuristique II

- la procédure séquentielle [A] sera bornée par $n^3/6$ (cas où l'on aboutit à une classe unique);
- la procédure (elle-même récursive) [T] est de l'ordre de Hn^2, H pouvant être supérieur à n;
- la procédure (récursive) [ES] consiste à explorer parmi toutes les cases extérieures aux classes formées, celles de valeur c_{ij} les plus négatives pouvant servir de "pivot". L'algorithme est donc approximativement de l'ordre de Kn^3, K étant une constante dépendant des données.

La convergence de l'algorithme est évidente. En effet, le critère d'arrêt étant défini par:
On arrête l'algorithme dès que la séquence : procédure [A] non améliorante - procédure [T] non améliorante - procédure [ES] non améliorante se produit au cours de l'algorithme.
Supposons en effet que ce critère d'arrêt ne puisse jamais s'appliquer, ceci signifierait que dans la séquence [A] → [T] → [ES] au moins une des procédures serait à chaque fois améliorante, l'amélioration portant sur une valeur entière du type (2 K) (K entier positif), la fonction F(Y) tendrait donc vers + ∞, or F(Y) est bornée supérieurement. CQFD.

APPLICATION ET EXEMPLES

Dans les deux petits exemples qui suivent je montre comment l'addition de la procédure ES correspondant à la Condition 4 permettra d'une part (exemple 1) d'améliorer considérablement la valeur de F(Y), d'autre part (exemple 2) de débloquer la procédure des Transferts et de pouvoir atteindre l'optimum global.

EXEMPLE 1

Considérons le tableau de comparaisons (6X6) suivant construit à partir des opinions de similarités de 9 "juges" sur 6 objets.

	1	2	3	4	5	6
1		8	5	7	7	2
2	8		2	5	1	7
3	5	2		8	7	1
4	7	5	8		3	2
5	7	1	7	3		6
6	2	7	1	2	6	

Le tableau C' associé soit C' = 2C - 9 (c'_{ii} =0) est donné par:

	1	2	3	4	5	6
1	0	7	1	5	5	-5
2	7	0	-5	1	-7	5
3	1	-5	0	7	5	-7
4	5	1	7	0	-3	-5
5	5	-7	5	-3	0	3
6	-5	5	-7	-5	3	0

L'application de l'algorithme récursif à 2 procédures :

1 [A] --► 2 [T] --► 1 [A] --► etc.. portant sur les seules procédures d'agglomération et de transfert conduit à la partition solution Yo définie par:

$$Yo \begin{cases} \text{Classe 1} = 1, 2, 3, 4 = \mathcal{C}_1 \\ \text{Classe 2} = 5, 6 = \mathcal{C}_2 \end{cases}$$

La valeur de F(Y) est égale à F(Yo) = 38

On voit sur la matrice C' précédente que pour Yo:

i) Les conditions de "non agglomération" sont vérifiées car

$$\Delta(\mathcal{C}_1, \mathcal{C}_2) = 2\left(\sum_{i\in\mathcal{C}_1}\sum_{j\in\mathcal{C}_2} c'_{ij}\right) = -24$$

d'où $\Delta(\mathcal{C}_1, \mathcal{C}_2) < 0$

ii) Les conditions de "non-transferts" sont vérifiées en effet:

$C_1(1) = 26$ et $C_1(2) = 0$
$C_2(1) = 6$ et $C_2(2) = -4$
$C_3(1) = 6$ et $C_3(2) = -4$
$C_4(1) = 26$ et $C_4(2) = -18$

d'où $C_i(1) > C_i(2)$ pour $i \in \mathcal{C}_1$
et de même
$C_5(1) = 0$ et $C_5(2) = 6$
$C_6(2) = -24$ et $C_6(2) = 6$

d'où $C_i(1) < C_i(2)$ pour $i \in \mathcal{C}_2$

donc la partition Yo est minimum local non améliorable par l'algorithme : [A] → [T] → [A] ...

Par contre si nous transferons simultanément:
2 de la classe $\mathcal{C}_1$ à la classe $\mathcal{C}_2$
et
5 de la classe $\mathcal{C}_2$ à la classe $\mathcal{C}_1$

Le "delta" de ce double transfert simultané est égal à:

$$\Delta(2,5) = C_2(2) - C_2(1) + C_5(1) - C_5(2) - 4\,c'_{25}$$
$$\Delta(2,5) = -16 + 28 \cong 12$$

En posant: Y1 $\begin{cases} \text{Classe 1} = 1,5,3,4 = C_1 - \{2\} + \{5\} \\ \text{Classe 2} = 2,6 \quad\quad = C_2 - \{5\} + \{2\} \end{cases}$

On a $F(Y1) = 38 + 12 = 50$

On a donc amélioré ici la fonction F(Y) de près de 25 % comme le montre la matrice C' (permutée suivant 1 5 3 4 2 6)

	1	5	3	4	2	6
1	0	5	1	5	7	-5
5	5	0	5	-3	-7	3
3	1	5	0	7	-5	-7
4	5	-3	7	0	1	-5
2	7	-7	-5	1	0	5
6	-5	3	-7	-5	5	0

Conclusion: le rajout de la procédure des "doubles échanges simultanés" permet ici sous la forme [A]→[T]→[ES]→etc.. d'obtenir la solution optimale.

EXEMPLE 2

Considérons maintenant un exemple (6X6), défini par le tableau collectif de similarités de 9 "juges" suivant:

	1	2	3	4	5	6
1		9	4	9	0	0
2	9		6	0	8	5
3	4	6		7	4	5
4	9	0	7		7	4
5	0	8	4	7		6
6	0	5	5	4	6	

Le tableau C' associé C'=2C - 9 est donné par:

	1	2	3	4	5	6
1	0	9	-1	9	-9	-9
2	9	0	3	-9	7	1
3	-1	3	0	5	-1	1
4	9	-9	5	0	5	-1
5	-9	7	-1	5	0	3
6	-9	1	1	-1	3	0

L'application de l'algorithme à deux procédures [A]→[T]→[A] etc.. donne la solution suivante:

$$Yo \begin{cases} \text{Classe 1: } 1,2 = \mathcal{C}_1 \\ \text{Classe 2: } 3,4,5,6 = \mathcal{C}_2 \end{cases}$$

et F(Yo)=42, Yo est un optimum local car les conditions de "non transfert" et de "non agglomération" sont vérifiées.

Si l'on rajoute la procédure [ES] à l'algorithme précédent il devient:

[A] → [T] → [ES] → [A] -- etc..

A partir de Yo, il convient donc de chercher les pivots les plus négatifs (une règle empirique consiste à ne tester que les couples (i,j) pour lesquels $c'_{ij} < 0$ et $|c'_{ij}| > \underset{i,j}{\text{Max}}(|c'_{ij}|/2)$).

Dans le cas présent il convient de calculer $\Delta(1,5)$, $\Delta(1,6)$ et $\Delta(2,4)$ on obtient:

$\Delta(1,5) = -16$
$\Delta(1,6) = -24$
$\Delta(2,4) = 4$

Donc comme le seul "double échange" avantageux est (2,4) , on transfère simultanément: 2 de $\mathcal{C}_1$ à $\mathcal{C}_2$ et 4 de $\mathcal{C}_2$ à $\mathcal{C}_1$.

On obtient alors la partition Y1 avec

$$Y1 \begin{cases} \text{Classe 1 : } 1,4 = \mathcal{C}_1 \\ \text{Classe 2 : } 3,2,5,6 = \mathcal{C}_2 \end{cases}$$

et F(Y1) = F(Yo) + 4 = 46

La nouvelle matrice C' permutée (par rapport à 1 4 3 2 5 6) est donnée par:

	1	4	3	2	5	6
1	0	9	-1	9	-9	-9
4	9	0	5	-9	5	-1
3	-1	5	0	3	-1	1
2	9	-9	3	0	7	1
5	-9	5	-1	7	0	3
6	-9	-1	1	1	3	0

Mais on constate à partir de Y1 que si les conditions de "non agglomération" sont vérifiées puisque $\Delta(\mathcal{C}_1, \mathcal{C}_2) = -20$, les conditions de "non transferts" ne sont pas toutes vérifiées, en effet:

$C_3(1) = 8$ et $C_3(2) = 6$

On peut donc transférer 3 de sa classe $\mathcal{C}_2$ dans la classe $\mathcal{C}_1$ et l'on obtient:

$$Y2 \quad \begin{cases} \text{Classe 1: } 1,4,3 = \mathcal{C}_1 + \{3\} \\ \text{Classe 2: } 2,5,6 = \mathcal{C}_2 - \{3\} \end{cases}$$

et $F(Y2) = F(Y1) + C_3(1) - C_3(2) = 46 + 2 = 48$

Aucune procédures [A] [T] ou [ES] n'est améliorante à partir de Y2 donc Y2 est un optimum pour l'algorithme [A] → [T] → [ES] . De plus cet optimum a priori "local" est aussi optimum "global".

En conclusion: on voit sur cet exemple que non seulement la procédure [ES] est améliorante mais qu'elle permet le "déblocage" de la procédure [T].

ANNEXE 2

(PROPRIETE PARETIENNE)

Agrégation de tableaux individuels C^k transitifs par un tableau collectif Y transitif et symétrique (équivalence).

On a le résultat suivant:
Si j est en relation avec i pour l'ensemble des m juges et si i est en relation avec j pour l'ensemble des m juges, alors j est en relation avec i dans le solution collective optimale; c'est à dire j et i sont dans la même classe dans la solution optimale.

Appelons encore Y^1 la solution optimale. Ce résultat s'exprime par:

$$c^1_{ij} = m \text{ et } c^1_{ji} = m \text{ impliquent } y^1_{ij} = 1$$

Démonstration:

Supposons que $c^1_{ij} = m$ et $c^1_{ji} = m$ mais que $y^1_{ij} = 0$. Comme Y^1 est symétrique $y^1_{ji} = 0$ c'est à dire i et j ne sont pas dans la même classe.

Appelons:
K_i les objets qui sont dans la même classe que i (i excepté)
K_j les objets qui sont dans la même classe que j (j excepté)
(Ces ensembles peuvent très bien être vides).

La valeur de la fonction économique est donnée par :

$$F(C,Y^1) = Cte + \sum_{k \in K_i} (c^1_{ik} + c^1_{ki})$$

Les termes contenant les indices i sont indiqués. La partie constante contient tous les termes où l'indice i ne figure pas.

Appelons Y^2 la solution collective dans laquelle i est mis dans la même classe que j, la valeur correspondante de la fonction économique est donnée par:

$$F(C,Y^2) = Cte + \sum_{k \in K_j} (c^1_{ik} + c^1_{ki}) + c^1_{ij} + c^1_{ji}$$

La constante étant la même que précédemment.

Posons:

$$S_i = \sum_{k \in K_i} (c^1_{ik} + c^1_{ki}) \quad (\text{Si } K_i = \emptyset,\ S_i = 0)$$

et $$S_j = \sum_{k \in K_j} (c^1_{ik} + c^1_{ki}) \quad (\text{Si } K_j = \emptyset,\ S_j = 0)$$

a) $S_j \geq S_i$

Dans ce cas $F(C,Y2) - F(C,Y^1) = S_j - S_i + c^1_{ij} + c^1_{ji} \geq 2m$ et Y^1 n'est pas solution optimale.

b) $S_i > S_j$

Dans ce cas écrivons la fonction économique associée à Y^1 sous la forme suivante :

$$F(C,Y^1) = Cte + \sum_{k\in K_j} (c'_{jk} + c'_{kj})$$

Les termes contenant les indices j, cette fois, sont indiqués. La partie constante contient tous les termes où l'indice j ne figure pas

Appelons Y^3 la solution collective dans laquelle j est mis dans la même classe que i, la fonction économique associée est donnée par:

$$F(C,Y^3) = Cte + \sum_{k\in K_i}(c'_{jk} + c'_{kj}) + c'_{ij} + c'_{ji}$$

La constante étant la même que dans la formule précédente.

Comme par hypothèse chaque tableau C'^k est transitif le tableau collectif C' verifie :

$c'_{ij} + c'_{jk} - m \leq c'_{ik}$ d'une part et $c'_{ji} + c'_{ik} - m \leq c'_{jk}$ d'autre part

Comme par hypothèse $c'_{ij} = m$ et $c'_{ji} = m$, on a respectivement:

$c'_{jk} \leq c'_{ik}$ et $c'_{ik} \leq c'_{jk}$ d'où $c'_{ik} = c'_{jk}$

De même de:

$c'_{kj} + c'_{ji} - m \leq c'_{ki}$ et $c'_{ki} + c'_{ij} - m \leq c'_{kj}$ on en déduit

$$c'_{ki} = c'_{kj}$$

Il en résulte que :

$$\sum_{k\in K_j}(c'_{jk} + c'_{kj}) = \sum_{k\in K_j}(c'_{ik} + c'_{ki}) = S_j$$

$$\sum_{k\in K_i}(c'_{jk} + c'_{kj}) = \sum_{k\in K_i}(c'_{ik} + c'_{ki}) = S_i$$

On a donc $F(C,Y^3) - F(C,Y^1) = S_i - S_j + c'_{ij} + c'_{ji} \geq 2m$

puisque $S_i - S_j > 0$ par hypothèse.

Dans ce cas également Y^1 n'est pas non plus solution optimale. D'où le résultat.

En résumé Y^1 ne peut être solution optimale, ce qui contredit l'hypothèse $y'_{ij} = 0$; d'où le résultat annoncé.

Remarque:
Les résultats présentés ici ne font aucune hypothèse, excepté la transitivité, quant à la nature des tableaux de données individuels.

ANNEXE 3

(INDICE DE CONCORDANCE GENERALISÉ)

1) On a vu en (T1 p.44) l'expression d'un indice de "concordance" généralisé, proposé pour mesurer la qualité des données vis à vis de l'hypothèse "d'unanimité" des "juges" ou "critères".
Cet indice est donné par:

$$I(conc) = \frac{\sum_{i \neq j} c'^{2}_{ij}}{m^2 \; n(n-1)} \qquad (1)$$

avec m = nombre de juges
n = nombre d'objets.

Cet indice varie de 0 à 1 pour m pair,
de 1/m2 à 1 pour m impair.

On peut d'ailleurs (Michaud) définir un indice indépendant de la parité de m en posant:

$$I'(conc) = 1 - \frac{\sum_{i \neq j} c_{ij}(m-c_{ij})}{[m/2][(m+1)/2] \; n(n-1)} \qquad (2)$$

avec [] = partie entière.

Cet indice varie de 0 à 1, quelque soit m.

Le problème dont il sera question dans cette annexe est la généralisation à l'Agrégation des Similarités d'un problème étudié pour l'Agrégation des Préférences (Marcotorchino, (R12) pp. 70 à 90), et qui consiste à chercher la "partition Y" supplémentaire émise par un "juge" ou "critère" virtuel J qui maximiserait la concordance entre son opinion et les opinions ou les similarités émises par les m "juges" de départ.
Si l'on considère pour une plus grande simplicité des calculs l'indice défini par la formule (1), chercher la partition Y qui maximise la "concordance" mutuelle entre les opinions des m "juges" ou "critères" de départ et l'opinion du juge virtuel supplémentaire donné par Y, revient à maximiser la fonction de Y suivante:

$$I(conc,Y) = \frac{\sum_{i \neq j} (c'_{ij} + 2y_{ij} - 1)^2}{(m+1)^2 \; n(n-1)}$$

pour $Y \in Pn$ (ensemble des partitions de n objets).

L'expression de I(conc,Y) se simplifie sous la forme:

$$I(conc,Y) = \frac{4 \sum_{i \neq j} c'_{ij} y_{ij} + \sum_{i \neq j} (c'_{ij} - 1)^2}{(m+1)^2 \quad n(n-1)}$$

Donc,

$\underset{Y \in Pn}{\text{Max}}$ I(conc,Y) est équivalent à

$$\underset{Y \in Pn}{\text{Max}} \sum_{i \neq j} c'_{ij} y_{ij}$$

En d'autres termes, maximiser marginalement la concordance I(conc,Y) pour $Y \in Pn$, revient à maximiser F(Y) (fonction économique de la Règle de Condorcet sous les contraintes $Y \in Pn$); c'est à dire à résoudre le problème d'optimisation du modèle b) ((T1) p.16).

Remarque: cette propriété est également vérifiée pour un problème d'agrégation des préférences. Ceci montre le caractère général de l'indice ainsi proposé.

2) Si l'on utilise le "coefficient of agreement" de KENDALL "U" que nous avons noté "Wo" dans ((R12) p.86):

$$U(m) = \frac{4 \sum_{i \neq j} c_{ij} (c_{ij} - 1)}{m(m-1) \quad n(n-1)} - 1$$

Nous avons vu en ((R1) p.87) que la propriété précédente était vérifiée pour cet indice dans le cas où le juge supplémentaire Y donne une opinion Y représentant un ordre total.
Dans le cas de l'agrégation de similarités, l'inconnue Y doit représenter une partition et comme:

$$U(m+1,Y) = \frac{4 \sum_{i \neq j} (c_{ij} + y_{ij})(c_{ij} + y_{ij} - 1)}{(m+1)m \quad n(n-1)} - 1$$

$$U(m+1,Y) = \frac{4 \sum_{i \neq j} c_{ij}^2 + 8 \sum_{i \neq j} c_{ij} y_{ij} - 4 \sum_{i \neq j} c_{ij}}{(m+1)m \quad n(n-1)}$$

donc:

$\underset{Y \in Pn}{\text{Max}}$ U(m+1,Y) est équivalent à maximiser

$\sum_{i \neq j} c_{ij} y_{ij}$ pour $Y \in Pn$,

or on a vu en ((T1) p.18) que la maximisation de cette fonction conduit à la solution triviale $y_{ij} = 1$ quelque soit (i,j), c'est à dire à la solution où tous les objets sont situés dans la même classe (partition grossière).
Ceci montre l'impossibilité d'utiliser cet indice pour mesurer la notion de concordance dans les problèmes d'agrégation des similarités.

L'indice "U" de Kendall n'a donc pas un caractère général.

En d'autres termes, on dira qu'un indice de concordance est consistant par rapport à une règle d'agrégation lors de l'adjonction d'un "juge virtuel" dont le tableau de comparaisons est représenté par Y, si:

$$\max_{Y \in R} I(conc,Y) \text{ "est équivalent" à } \max_{Y \in R} F(Y)$$

F étant le critère caractérisant la règle d'agrégation;
R le type de la relation Y.

3) A titre d'illustration considérons l'indice de concordance modifié (Marcotorchino) en cas de non réponses en Agrégation des Préférences ((T1) page 45), et verifions si cet indice de concordance est "consistant" par rapport aux règles d'agrégation que nous avons utilisées précédemment.

En effet cet indice est défini par:

$$I(conc) = 1 - 4 \sum_{i \neq j} \frac{c_{ij} c_{ji}}{(c_{ij}+c_{ji})^2 n(n-1)}$$

Si l'on rajoute une opinion supplémentaire Y représentant un ordre total on a: $y_{ij} + y_{ji} = 1$ (3)

donc l'indice devient:

$$I(conc) = 1 - 4 \sum_{i \neq j} \frac{(c_{ij}+y_{ij})(c_{ji}+y_{ji})}{(c_{ij}+y_{ij}+c_{ji}+y_{ji})^2 n(n-1)}$$

A priori ce coefficient est non linéaire en Y à cause du dénominateur, mais du fait de la propriété (3) la variable y_{ij} disparait au dénominateur qui devient:

$$(c_{ij}+c_{ji}+1)^2 \, n(n-1)$$

le numérateur etant égal quant à lui à:

$$4 c_{ij} c_{ji} - 4 (c_{ij}-c_{ji}) y_{ij} + 4 c_{ij}$$

Donc $\max_{Y \in R} I(conc,Y)$ revient à maximiser la fonction de Y:
R (relation d'ordre total)

$$F'(Y) = \sum_{i \neq j} \frac{(c_{ij} - c_{ji})}{(c_{ij} + c_{ji} + 1)^2} y_{ij} \qquad (4)$$

Maximiser cette expression n'est pas équivalent à maximiser $F(Y) = \Sigma\, c_{ij}\, y_{ij}$ (règle de Condorcet) dans le cas où $Y \in R$, R (relation d'ordre total).

I(conc) n'est pas "consistant" avec la règle de Condorcet mais est consistant avec une règle de type "majorité relative par paires" définie par la fonction économique linéaire en y_{ij}, F', donnée par la formule (4).

LISTE DES TEXTES PRESENTES

- T1 : AGRÉGATION DES SIMILARITÉS EN CLASSIFICATION AUTOMATIQUE,

- T2 : OPTIMISATION EN ANALYSE DES DONNÉES RELATIONNELLES

- T3 : OPTIMIZATION IN EXPLORATORY DATA ANALYSIS

T1

AGREGATION DE SIMILARITES EN CLASSIFICATION AUTOMATIQUE

F. Marcotorchino P. Michaud

Centre Scientifique IBM France

I- INTRODUCTION

Cet article a pour but la présentation, par une approche tout à fait différente de celles habituellement utilisées par les spécialistes de l'Analyse des Données, d'une méthodologie générale, utilisable pratiquement pour résoudre divers types de problèmes d'analyse des données (en particulier certains problèmes de classification automatique).

La méthode proposée est une méthode d'AGREGATION DE DONNEES, ayant pour but la recherche d'une "relation particulière" sur n objets, soit un ordre total (dans le cas de l'agrégation des préférences), soit une partition (dans les cas de l'agrégation de similarités ou agrégation classificatoire), soit encore d'autres relations plus complexes, "approximant au mieux" un ensemble de relations individuelles de départ représentant les données de m "critères" ou "juges" relatives aux n objets à analyser.

L'originalité de notre approche repose essentiellement sur l'utilisation de modèles linéaires généraux unifiant des problèmes a priori aussi différents que la recherche d'un "ordre" sur une population dans le cadre de l'agrégation des préférences ou que la recherche d'une "partition" de cette même population dans le contexte de la classification automatique.

Cette approche a été présentée la première fois dans (18) et dans (16).

A titre d'exemple donnons ici quelques points qui différencient cette approche et les méthodes existantes dans le domaine sur lequel nous avons choisi d'insister ici:

c'est à dire la Classification Automatique.

Rappelons qu'en classification automatique, les méthodes existantes se divisent approximativement en deux grands types:
-les méthodes de classification hiérarchique ascendante;
-les méthodes de classification non hiérarchique.

Pour les premières, le résultat est en général un "arbre de classification" (on n'obtient pas de partition de la population), obtenu pas à pas sur des critères de "distances" minimales entre classes par agglomérations successives des classes (5) (12).

Pour les secondes, le résultat est une "partition" de la population obtenue (après fixation a priori du nombre de classes) par diminution d'un critère lié à des opérations de "recentrage - réaffectation" (9).

L'approche proposée ici diffère des deux précédentes tant du point de vue du résultat obtenu que du critère retenu:
-le résultat est certes une "partition", mais sans fixation a priori du nombre de classes;
-le critère repose sur l'utilisation d'une régle majoritaire ancienne puisqu'il s'agit de la fameuse "régle de Condorcet".

II- HISTORIQUE ET POSITION DU PROBLEME

Donnons tout d'abord la définition des termes de base de ce texte: "Agrégation des préférences" et " Agrégation des similarités". On appellera par la suite dans le texte :

Problème d'Agrégation des Préférences;
un problème où les données seront la représentation de préférences sur les objets (ordres, préordres, comparaisons préférentielles par paires, etc...) et où le résultat cherché, sera un Ordre Total.

de même on appellera:

Problème d'Agrégation de Similarités:
un problème où les données seront la représentation de similarités sur des objets (partitions , similarités par paires etc..) et où surtout le résultat cherché sera une Partition.

L'approche que nous allons proposer, dont le but est la résolution des deux problèmes précédents, bien que récente

de par les techniques employées est fondée sur des concepts anciens introduits par A. de CONDORCET dès 1785 (recherche d'une relation d'ordre collectif). Ce dernier avait déjà mis en évidence à l'époque l'aspect paradoxal de la recherche d'un ordre collectif à partir d'opinions individuelles (Effet Condorcet) et il fallut attendre 1950 pour voir s'édifier avec K. ARROW (1) une axiomatique de l'agrégation des opinions individuelles sous la forme d'un théorème d'impossibilité (4 axiomes: Non dictature, Unanimité, Indépendance, et Universalité). Aucune procédure ne vérifie simultanément les 4 axiomes précédents (sauf justement sous certaines conditions la règle majoritaire de Condorcet).

Cet aspect paradoxal se retrouve également dans le cas de l'agrégation de similarités, où le problème consiste alors à trouver une "relation d'équivalence collective" (partition des n objets) la plus "proche possible" de m partitions ou relations de similarités individuelles données au départ par m "juges" ou "critères" sur les n objets à analyser. Le paradoxe ne s'appelle plus dans ce cas "Effet Condorcet", mais "Effet Poincaré"

Comme nous venons de le dire les méthodes que nous allons présenter permettent de résoudre aussi bien les problèmes d'agrégation des préférences que d'agrégation des similarités. Or si, comme nous le montrerons plus loin les modèles associés à ces deux problèmes de nature a priori différente, ne diffèrent que très légèrement, il semble curieux de constater qu'historiquement ces problèmes ont été étudiés de façon tout à fait séparée et sans aucune connexion apparente.

En effet, alors que l'on peut faire remonter les premières idées de base sur "l'agrégation des préférences" à A. de Condorcet (1785), le problème de la recherche de "partitions dites centrales" (cas particulier de l'agrégation de similarités) n'a été posé pour la première fois par S. Regnier qu'en 1965 (24).
Quant au développement des connaissances sur ces deux problèmes, tant au niveau théorique et axiomatique qu'au niveau des algorithmes de résolution, l'état d'avancement était loin d'être identique jusqu'à aujourd'hui.

A l'étude du problème de "l'agrégation des préférences", outre Condorcet et Arrow déjà cités, on peut associer les noms de M.G. Kendall (14), G. Kemeny (13) qui ont posé le problème sous forme "metrique" dans le cas le plus simple de l'agrégation d'ordres totaux sans pour cela proposer de méthodes de résolution, et de D. Black (6) qui en a poursuivi l'étude axiomatique (à propos duquel nous aurons l'occasion de revenir ultérieurement). M.Barbut, dès 1966 (2), avait également souligné l'aspect métrique du problème d'agrégation.

En ce qui concerne "l'agrégation des similarités" dont la recherche d'une "partition centrale" n'est qu'un cas particulier (consistant à chercher la partition agrégeant des données elles-mêmes sous forme de partitions), à S.Regnier déjà cité, on peut associer les noms de B. Mirkin (22) et I.C. Lerman (15). Sous sa forme initiale le problème de la recherche d'une partition centrale n'avait pas été rattaché à l'approche générale.

Plus récemment J.P. Barthelemy (3) (4) et B.Monjardet (23) ontproposé la métrique précédente pour la recherche de relations collectives plus complexes que des ordres totaux ou des "partitions" (en particulier préordres, tournois, etc...), posant ainsi le problème général de la recherche de "Relations Centrales".

En ce qui concerne les méthodes de résolution, la distinction à faire entre "agrégation des préférences" et "agrégation des similarités" était encore plus nette.
En effet:

Les méthodes de résolution exactes associées à ce type d'agrégation des préférences" étaient toutes fondées,jusqu'à présent, sur le principe d'énumération implicite "branch and bound", (nous avons donné en (16) une liste de certains de ces algorithmes); mais en pratique il était impossible de les utiliser pour un nombre d'objets à classer supérieur à 20.
Seul J. de Cani (7) avait vu l'intérêt de l'utilisation de la programmation linéaire pour résoudre un problème voisin de celui dont nous traitons ici, après quelques essais infructueux limités à l'agrégation d'ordres totaux de taille très faible n=4, il a de lui même renoncé à cette approche et a proposé à son tour ultérieurement (8) un algorithme d'énumération implicite.

En ce qui concerne les méthodes de résolution associées à "l'agrégation des similarités", exceptée l'heuristique de S.Regnier (24) pour les "partitions centrales", aucun algorithme a fortiori exact n'avait été proposé jusqu'à ce jour.

Aucun des auteurs précités n'avait, tout au moins à notre connaissance, fait le lien entre l'agrégation des préférences et l'agrégation des similarités au niveau des méthodes de résolution, et proposé un modèle général.

A ce niveau notre contribution a permis de synthétiser et d'améliorer cette situation en proposant l'approche méthodologique suivante:
i) Construction d'un modèle linéaire général de l'agrégation de relations de tous types.
ii) Création de méthodes de résolution exactes et approchées adaptées aux deux types de problèmes d'agrégation

considérés ici.

En effet nous avons modélisé pour la première fois (18) ces divers problèmes de recherche de relations centrales sous une forme linéaire générale permettant ainsi de développer des méthodes exactes de résolution en des temps de calcul raisonnables, même pour des tailles de n relativement importantes ($n \leq 80$).
Enfin, nous avons montré que l'ensemble de ces problèmes pouvait être considéré comme l'utilisation de la règle de la majorité de Condorcet (fonction critère linéaire), sous des contraintes linéaires dépendant de la nature de la relation cherchée (17) (19).

En résumé, le problème général que nous allons traiter consiste à rechercher une relation collective (ordre total ou partition entre autre) "approximant au mieux" un ensemble de relations individuelles de départ.

La modélisation et la résolution du problème repose sur les 5 concepts fondamentaux suivants:

1) Présentation des données sous forme relationnelle. (On montrera que bon nombre de données usuelles se présentent sous forme de relations).
2) Utilisation d'un critère linéaire d'agrégation qui n'est autre que la règle majoritaire de Condorcet, dont on montrera qu'elle est équivalente à une approche métrique;
3) Linéarisation des contraintes portant sur la relation cherchée (la relation cherchée que nous appelerons Y par la suite vérifiera un certain nombre de contraintes linéaires);
4) Utilisation de la Programmation Linéaire pour la résolution des problèmes ainsi posés (en particulier utilisation du système MPSX/370);
5) Utilisation d'indices de validation du problème, tant au niveau des données que des résultats; ceci afin de répondre aux questions fondamentales suivantes:
- est-il raisonnable de vouloir classer ou classifier les données de départ?
- le résultat obtenu est-il significatif?

III- DEFINITION DU PROBLEME GENERAL DE L'AGREGATION RELATIONNELLE

3-1) PRESENTATION GENERALE DES DONNEES

Afin de définir la méthode proposée, nous supposons ici que les données se présentent déjà sous forme relationnelle. Nous verrons dans un paragraphe ultérieur comment la prise

en compte des données usuelles les plus variées peut se faire sous cette forme.

On a donc au départ m relations individuelles correspondant soit aux opinions de m juges soit aux mesures faites sur les n objets pour m "critères".

a) Tableaux de données individuels

Chaque relation initiale, correspondant aux données individuelles du "juge" ou "critère" numero k sera représentée par un tableau de "comparaisons par paires" noté C^k. Ce tableau de comparaisons est un tableau carré (nXn) à valeurs (0 ou 1) défini comme suit:

$c_{ij}^k = 1$ Si i "est en relation avec " j pour le "juge" k

$c_{ij}^k = 0$ Sinon.

Par convention les valeurs c_{ii}^k du tableau ne seront pas considérées

Compte tenu du problème à traiter et de la nature des données, la signification des valeurs c_{ij}^k sera différente. Par exemple:
- i "est en relation avec" j signifie: i"est préféré à" j si C^k est un tableau de données de préférences.
- i "est en relation avec" j signifie: i "est semblable à" j si C^k est un tableau de similarités.

Remarque:

Chaque tableau C^k peut représenter n'importe quelle relation (en particulier des ordres totaux, des préordres totaux, des quasi ordres, des tournois , des relations symétriques, des relations d'équivalence etc..). (Voir détails en (20))

A chaque relation binaire représentée par un tableau C^k, on peut associer un autre tableau de comparaisons, noté C'^k, ce nouveau tableau est un tableau carré (nXn) à valeurs (-1,+1), défini par:

$c'^k_{ij} = 1$ Si i "est en relation avec" j pour le "juge" ou "critère" k

$c'^k_{ij} = -1$ Sinon.

Les valeurs c'^k_{ij} peuvent être exprimées au moyen des c_{ij}^k par:

$$c'^k_{ij} = 2\, c_{ij}^k - 1$$

Comme nous le verrons au Chapitre VII les tableaux C'^k et C^k peuvent être obtenus directement ou indirectement par l'intermédiaire de tableaux annexes (tableaux de rangs de notes, de ditances etc...).

b) Tableaux collectifs

En sommant tous les tableaux de comparaisons individuels C^k et C'^k, on obtient des tableaux de données collectifs notés:

$$C = \sum_k C^k \quad \text{et} \quad C' = \sum_k C'^k$$

$$C' = 2C - m$$

Le terme (i,j) du tableau C est donné par:

$$c_{ij} = \sum_k c^k_{ij}$$

Le terme (i,j) du tableau C' est donné par:

$$c'_{ij} = \sum_k c'^k_{ij}$$

c'ij = Nombre de "juges" ou "critères" pour lesquels i "est en relation avec" j moins Nombre de "juges" ou "critères" pour lesquels i "n'est pas en relation avec" j.

De ce fait c'ij > 0 signifie qu'une majorité de "juges" ou "critères" considèrent que i "est en relation avec" j.

Remarque:

On peut généraliser l'approche précédente au cas où les critères ou juges Ck sont pondérés par un coefficient $p_k \geq 0$. On a alors:

$$C = \sum_k p_k C^k \quad \text{et} \quad C' = \sum_k p_k C'^k$$

La relation liant C' à C est alors donnée par:

$$C' = 2C - \left(\sum_k p_k\right)$$

Dès lors tout ce qui sera dit par la suite sur la formulation la plus simple restera valable pour la formulation avec pondérations.

3-2) DEFINITION DE LA RELATION COLLECTIVE CHERCHEE Y.

Pour les deux problèmes (Agrégation des Préférences et de Similarités), nous recherchons une relation collective transitive (soit un ordre total , soit une relation

d'équivalence "partition") "approximant de la meilleure façon possible" les m relations binaires individuelles de départ C^k. Cette procédure de recherche d'une relation dite "centrale", conduira à l'obtention d'une relation Y.
Cette relation collective cherchée sera l'inconnue de notre modèle et sera représentée elle aussi par un tableau binaire carré (nXn) défini par:

$y_{ij} = 1$ Si i est "en relation avec" j dans le résultat final;

$y_{ij} = 0$ Sinon.

Comme pour les données les termes "est en relation avec" dépendent du contexte de l'agrégation que l'on désire effectuer et de la nature du résultat attendu.

Exemples:

- $y_{ij} = 1$ signifie: i est classé avant j si Y représente un ordre total;

- $y_{ij} = 1$ signifie: i est dans la même classe que j si Y est une partition.

En règle générale Y peut représenter n'importe quel autre type de relation binaire: quasi-ordre, semi-ordre, préordre total, etc... (18) et (20).

a) Récupération des résultats sous forme non matricielle

Supposons que nous ayons trouvé la ou les solutions optimales Y^* d'un problème d'agrégation des préférences ou d'agrégation des similarités alors, à partir des solutions optimales Y^* données sous forme de tableaux binaires, la présentation des classements ou des partitions associés se fait de la façon suivante:

1) Cas où Y^* est un ordre total:

Si Y^* est solution optimale d'un problème d'agrégation des préférences on déduit le classement optimal r_i^* de l'objet i de la façon suivante:

$$r_i^* = \sum_j y^*_{ji} + 1 = n - \sum_j y^*_{ij}$$

2) Cas où Y^* est une partition

Si Y^* est la solution optimale d'un problème d'agrégation des similarités on obtient la partition optimale en posant: $y^*_{ij}=1$ ==> i est dans la même classe que j.

b) Généralisation à d'autres types de solutions Y

Il est bien évident que l'Agrégation Relationnelle Générale permet des combinaisons entre le type de résultat cherché et le type de données à agréger autres que celles rencontrées en "agrégation des préférences" et en "agrégation des similarités".

Exemples:
Chercher Y (Partition) la plus proche possible de données C^k représentant des préordres, ou chercher la relation Y (Préordre Total) la plus proche possible de données représentant des quasi-ordres sont deux configurations tout à fait du ressort de l'agrégation relationnelle.

3-3) DEFINITION D'UN "CRITERE" D'AGREGATION

Nous avons maintenant à définir ce que nous entendons par les termes :"approximer de la meilleure façon possible". Pour cela considérons la fonction de Y suivante:

$$F(Y) = \sum_i \sum_j c'_{ij} y_{ij}$$

où c'_{ij}, comme nous l'avons déjà vu, représente le nombre de "juges" ou "critères" mettant i "en relation avec" j moins le nombre de "juges" ou "critères" ne mettant pas i "en relation avec" j.

La fonction F(Y) est alors trivialement maximum pour Y^c définie par:

$y^c_{ij} = 1$ Si $c'_{ij} > 0$

$y^c_{ij} = 0$ Si $c'_{ij} < 0$

$y^c_{ij} = 0$ ou 1 si $c'_{ij} = 0$

Comme nous l'avons déjà vu $c'_{ij} > 0$ implique qu'une majorité de "juges" a mis i "en relation avec" j; en conséquence Y. n'est rien d'autre que le résultat de l'application directe d'une règle majoritaire au tableau collectif de comparaisons C', c'est justement cette règle qu'avait proposée Condorcet dès 1785 pour l'agrégation d'ordres totaux.

Cependant ce critère peut s'interpréter également de façon "métrique".

3-4) INTERPRETATION METRIQUE DE LA FONCTION F(Y)

L'approche que nous proposons ici, fondée sur l'application de la règle de Condorcet, repose également sur une représentation "métrique". En effet la fonction économique

F(Y), présentée au paragraphe précédent, s'interprète également comme une fonction d'éloignement.

a) Définissons tout d'abord la distance $d(C^k,Y)$ entre la relation cherchée Y et un tableau relationnnel et individuel de données C^k.

Cette distance s'écrit

$$d(Y,C^k) = \sum_i \sum_j |y_{ij} - c^k_{ij}|$$

Cette distance se ramène à la forme linéaire en Y suivante: (18 p.27)

$$d(Y,C^k) = \sum_i \sum_j c^k_{ij} - \sum_i \sum_j c'^k_{ij}\, y_{ij}$$

Remarque:

Cette distance entre C^k et Y, que l'on aurait pu écrire également $d(C^k,Y) = \Sigma\Sigma\, |y_{ij} - c^k_{ij}|^m, m \geq 1$ puisque les valeurs y_{ij}, $c^k_{ij} \in (0-1)$, est en fait la distance de la "différence symétrique" entre les deux relations Y et C^k.

b) Afin de tenir compte de la notion "d'approximation" sous-jacente dans cette approche, on mesurera "l'éloignement" de Y à toutes les relations individuelles C^k par la fonction de Y suivante:

$$F'(Y) = \sum_k d(Y,C^k)$$

Cette expression est également linéaire en termes de yij et dépend seulement des valeurs c_{ij} et c'_{ij} des tableaux de données collectifs C et C' (18) (16).

$$F'(Y) = \sum_i \sum_j c_{ij} - \sum_i \sum_j c'_{ij} y_{ij}$$

Minimiser en Y cette fonction F'(Y) reviendra à chercher la relation Y la plus "proche possible" de l'ensemble de m relations binaires de départ.
Dès lors, il apparait la relation évidente entre les fonctions de Y, F(Y) et F'(Y) suivante:

$$F'(Y) = Cte - F(Y)$$

Et l'on obtient les deux résultats fondamentaux suivants:

i) Les fonctions objectif F(Y) et F'(Y) sont linéaires en y_{ij}

ii) Les problèmes:

$$\min_Y \; F'(Y) \quad \text{(présentation métrique)}$$

et

$$\max_Y \; F(Y) \quad \text{(présentation au moyen de la règle de Condorcet)}$$

sont deux problèmes équivalents.

En d'autres termes chercher une relation Y la plus "proche possible" des m relations de départ C^k , ou maximiser les majorités collectives émises par paires sont deux critères d'agrégation équivalents.

Comme nous l'avons déjà vu ces deux problèmes Min F'(Y) et Max F(Y)
sont trivialement résolus en appliquant la règle majoritaire de Condorcet.

Remarque:

Il faut noter qu'ici Y est solution d'une fonction économique sans contraintes et peut donc représenter n'importe quelle relation binaire. Bien entendu sauf si le tableau C' a une structure particulière, la solution Y. ne sera pas en général transitive. Or nous cherchons une relation qui devra être soit une relation d'ordre total soit une relation d'équivalence, c'est à dire transitive; la seule utilisation de la règle majoritaire de Condorcet ne nous garantit donc pas contre une solution dénuée de fondement (vis à vis du problème posé).

Ce fait est du en particulier à l'existence de "paradoxes" inhérents à ce type de problèmes.

L'un de ces paradoxes est le "fameux" paradoxe de Condorcet. Le principe de ces paradoxes repose sur le fait suivant: l'agrégation de données individuelles "transitives" peut donner un résultat Y "intransitif" si on applique la règle de la majorité.

En supposant que des "juges" ou "critères" aient donné des opinions individuelles transitives (ordres totaux ou partitions) sur 3 objets x,y et z, alors il y aura (vis à vis de la règle de Condorcet):

Effet Condorcet

si
Une majorité de "juges" ayant préféré collectivement x à y : $(x > y)$, une majorité de "juges" ayant également préféré collectivement y à z $(y > z)$, on a une majorité de "juges" qui préfère collectivement z à x $(z > x)$, au lieu de la transitivité attendue ($(x > z)$, x préféré à z par une majorité de juges).

De même il y aura:

Effet Poincare

si, sachant que pour une majorité de "critères" x est considéré comme similaire à y ($x \simeq y$), y comme similaire à z ($y \simeq z$), mais que x n'est pas considéré comme similaire à z par une majorité de "critères" ($x \not\simeq z$); au lieu de la transitivité attendue ($x \simeq z$) x considéré comme similaire à z par une majorité de "critères".

Comme l'application directe (fort simple) de la règle de Condorcet ne nous permet pas d'obtenir un résultat transitif à coup sûr, la relation Y cherchée doit donc vérifier un certains nombre de contraintes pour nous garantir l'obtention soit d'un "ordre total", soit d'une "partition", relations auxquelles nous nous sommes volontairement limités dans cet article. La vérification de ces contraintes sur Y va rendre les problèmes précédents "combinatoires", c'est à dire débouchant sur des calculs inextricables si l'on ne possède pas d'algorithmes performants.
- En effet dans le cas de l'agrégation des préférences il faudrait choisir parmi les n! ordres totaux Y celui ou ceux qui maximisent F(Y).
- Dans le cas de l'agrégation des similarités c'est parmi les Bn partitions de n objets (nombre de Bell) qu'il faudrait choisir la ou les partitions Y maximisant F(Y).

A titre d'illustration disons que $(17)! = 3{,}5 \times 10^{12}$ et $(67)! > 10^{100}$ et que $B(19) = 5{,}8 \times 10^{12}$.

Mais là encore (nous allons le voir au Chapitre suivant), ces contraintes s'exprimeront linéairement en fonction des valeurs y_{ij}; ceci nous permettra d'éviter l'énumération de toutes les solutions Y possibles (procédé impraticable pour $n \geq 15$).

Cependant bien qu'exceptionnels, il existe des cas où de par la structure particulière des données, la solution Y^c de Condorcet sera soit un ordre total, soit une relation d'équivalence, c'est à dire solution optimale du problème posé. Citons quelques unes de ces configurations exceptionnelles.

3-5) CAS OU LA REGLE DE CONDORCET SEULE DONNE UNE SOLUTION ADMISSIBLE

a) <u>Pour les problèmes d'agrégation des préférences:</u> (CONDITION DE BLACK)

Si les tableaux C^k, correspondant aux opinions des "juges" vérifient la condition de Black (6) (représentation unimodale du classement des n objets par chaque "juge", les objets étant rangés selon un ordre de référence), la solution Y^c obtenue par application directe de la règle de

Condorcet sur le tableau C' collectif sera un ordre total. Quoique rarement vérifiées en pratique, il arrive que ces conditions puissent l' être par des opinions individuelles quand un ordre de référence existe implicitement (par exemple en politique, ordre latent gauche - droite).

b) Pour les problèmes d'agrégation de similarités: (CONDITION D'EMBOITEMENT)
Si les tableaux C^k (tableaux individuels) représentent des partitions et si ces m partitions données sont emboitées c'est à dire que l'une est plus fine que l'autre (deux à deux), ce qui se traduit au niveau des tableaux de données C^k par : il existe un ordre de référence sur les partitions (1, 2, 3...ℓ,ℓ +1, ...m) tel que :

$$C^{\ell} \leq C^{\ell+1} \text{ ou } C'^{\ell} \leq C'^{\ell+1} \quad \forall\, \ell$$

alors toutes ces partitions peuvent être considérées comme les partitions compatibles d'une C.H.A (classification hierarchique ascendante) c'est à dire d'un arbre de classification; et dès lors, induire une similarité ultramétrique entre les n objets (12). De ce fait ceci entraine la relation suivante au niveau des tableaux collectifs C et C':

-1) $c_{ij} \geq \text{Min}\,(c_{ik}, c_{kj}) \quad \forall$ i,j,k tous différents

-2) $c'_{ij} \geq \text{Min}\,(c'_{ik}, c'_{kj}) \quad \forall$ i,j,k tous différents

Les conditions 1) ou 2) sont connues sous le nom de transitivité ultramétrique (20).

La solution Y^c obtenue par application de la règle de Condorcet au tableau collectif C' vérifiant la condition 2) sera une partition; en effet le tableau C' étant somme de tableaux symétriques sera symétrique, donc Y^c sera symétrique.

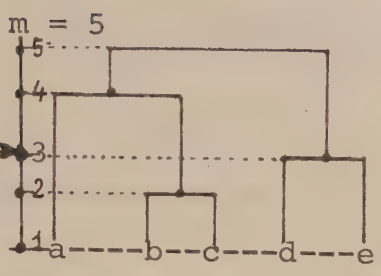
Si $c'_{ik} > 0$ et $c'_{kj} > 0$ alors $y_{ik} = 1$ et $y_{kj} = 1$ mais $c'_{ij} \geq \text{Min}(c'_{ik}, c'_{kj})$ implique $c'_{ij} > 0$ donc $y_{ij} = 1$, donc Y^c sera transitif.
Par suite Y^c symétrique et transitive représente une partition.

La solution optimale Y^* est même explicite dans le cas d'un nombre impair m de critères donnant des partitions emboitées, en effet

$$Y^c = C^{\frac{m+1}{2}} = Y^* \text{ optimale}$$

où $C^{\frac{m+1}{2}}$ est le tableau des données du "juge" ayant donné la "partition médiane" par rapport à l'échelle des emboitements.

Exemple:



Médiane de l'échelle des emboitements

P1= a, b, c, d, e
P2= a, (b,c), d, e
P3= a, (b,c), (d,e)
P4= (a,b,c), d, e
P5= (a,b,c,d,e)

Dans ce cas précis Y^C = P3 = a,(b, c),(d, e).

Remarque:

La "Condition d'Emboitement" pour l'agrégation des similarités est la parallèle à la "Condition de Black" pour les problèmes d'agrégation des préférences. Dans ce cas on ne suppose plus un ordre de référence mais une échelle de stratification de référence pour les partitions.

IV- LES MODELES LINEAIRES PROPOSES

4-1) LES CONTRAINTES LINEAIRES SUR LA RELATION CHERCHEE Y

-Si la relation cherchée Y est un ordre total, Y doit être (Transitive, Antisymétrique, Totale);
-Si la relation cherchée Y est une partition (relation d'équivalence), Y doit être: (Transitive, Symétrique).

Nous avons alors le résultat intéressant suivant:
Toutes ces propriétés peuvent être formulées comme des expressions linéaires des valeurs y_{ij} (voir (18) (20) pour plus de détails).

1) Transitivité: Elle s'exprime par les contraintes suivantes:

$$y_{ij} + y_{jk} - y_{ik} \leq 1 \quad \forall \quad (i,j,k) \text{ tous différents}$$

Ces inégalités expriment que si iRj et jRk alors iRk pour la relation Y.

2) Symétrie: Elle s'exprime par les contraintes suivantes:

$$y_{ij} - y_{ji} = 0 \quad \forall \ (i \neq j)$$

Cette contrainte signifie simplement que pour un problème de classification automatique si i est dans la même classe que j, alors j est dans la même classe que i.

3) Antisymétrie: Elle s'exprime par les contraintes suivantes:

$$y_{ij} + y_{ji} \leq 1 \quad \forall\ (i \neq j)$$

Ces inégalités signifient qu'au plus iRj ou jRi est vraie pour la relation Y.

4) Totalité: Elle s'exprime par les contraintes suivantes:

$$y_{ij} + y_{ji} \geq 1 \quad \forall\ (i \neq j)$$

Ces inégalités signifiant qu'au moins iRj ou jRi est vraie pour la relation Y.

Remarque:

Si Y est un ordre total (antisymétrique et total), Y doit vérifier simultanément les contraintes 3) et 4) ce qui conduit aux nouvelles relations linéaires en y_{ij}:

5) $$y_{ij} + y_{ji} = 1 \quad \forall\ (i \neq j)$$

Ces relations signifient que soit i est classé avant j soit j est classé avant i dans la relation cherchée Y.

De ce fait nous pouvons donner une interprétation générale des problèmes d'agrégation des préférences ou d'agrégation de similarités:

ces deux problèmes peuvent être considérés comme l'application de la règle majoritaire sous contraintes:
-contraintes 1) et 5) pour la recherche d'un ordre total;
-contraintes 1) et 2) pour la recherche d'une partition Y.

4-2) LES MODELES BIVALENTS

Il découle des résultats précédents que les problèmes d'agrégation des préférences et d'agrégation des similarités peuvent être formulés comme des problèmes de programmation linéaire en nombres entiers en variables bivalentes (0-1).

Les variables (0-1) sont les valeurs y_{ij} de la relation collective cherchée Y.

La fonction objectif à maximiser est la fonction F(Y) que nous avons déjà définie comme la formulation de la règle majoritaire de Condorcet; les contraintes sont les différentes propriétés que la relation Y doit vérifier.

Selon le type de relation Y cherchée nous obtenons donc les modèles linéaires suivants:

a) Modèle pour une relation d'ordre total

$$\text{Max} \sum_i \sum_{j \neq i} c'_{ij} y_{ij}$$

$$y_{ij} + y_{ji} = 1 \qquad i \neq j$$

$$y_{ij} + y_{jk} - y_{ik} \leq 1 \qquad i \neq j, j \neq k, i \neq k$$

$$y_{ij} = 0 \text{ ou } 1$$

b) Modèle pour une relation d'équivalence

$$\text{Max} \sum_i \sum_{j \neq i} c'_{ij} y_{ij}$$

$$y_{ij} - y_{ji} = 0 \qquad i \neq j$$

$$y_{ij} + y_{jk} - y_{ik} \leq 1 \qquad i \neq j, j \neq k, i \neq k$$

$$y_{ij} = 0 \text{ ou } 1$$

Il est tout à fait remarquable que des problèmes primitivement très différents quant au résultats cherchés: Recherche d'une agrégation des préférences (classement) et recherche d'une agrégation de similarités (classification), ne diffèrent du point de vue de la modélisation que par les équations $y_{ij}+y_{ji}=1$ et $y_{ij}-y_{ji}=0$

On trouvera dans (18) et plus spécialement dans (20) un certain nombre de modèles linéaires associés à des relations Y plus complexes.

Le nombre de contraintes de ces problèmes est de l'ordre $O(n^3)$, tandis que le nombre de variables est de l'ordre $O(n^2)$.

Remarque:

On peut réduire la taille de ces modèles en les transformant légérement (voir (16) p. 139, et (18) p. 45), mais l'ordre de grandeur reste le même.

4-3) QUELQUES VARIANTES SUR L'EXPRESSION DE F(Y)

a) Cas où Y est une relation d'ordre total

i) La maximisation de la fonction objectif F(Y) est équivalente, dans ce cas, à la maximisation du nombre d'accords définie par:

$$\sum_i \sum_j c_{ij} y_{ij}$$

La maximisation de F(Y) est aussi équivalente à la minimisation du nombre de désaccords défini par:

$$\sum_i \sum_j c_{ji}\, y_{ij}$$

Ce qui est également équivalent à la maximisation du nombre d'accords moins le nombre de désaccords.
Cette dernière expression:

$$\sum_i \sum_j (c_{ij} - c_{ji})\, y_{ij}$$

est elle même équivalente à la maximisation de la fonction de Y suivante, exprimée à l'aide du tableau C':

$$\sum_i \sum_j (c'_{ij} - c'_{ji})\, y_{ij}$$

Remarque:

Cette dernière formulation nous montre que l'on peut éliminer du calcul la partie symétrique des données, qui n'intervient pas dans la recherche d'une solution Y, devant par nature être antisymétrique.

En résumé disons que la fonction objectif F(Y), dans le cas particulier où l'on recherche une relation Y ordre total, conduit à plusieurs formulations équivalentes d'interprétation et de signification différentes.

ii) Lorsque les tableaux C^k ne représentent pas eux-mêmes des ordres totaux, le cas d'un tableau C^k unique est également intéressant. Si le tableau C est un tableau binaire quelconque, on se retrouve dans le cas d'un problème classique de théorie des graphes consistant à trouver un ordre total "approximant" au mieux une relation binaire quelconque. Un exemple particulier d'un tel problème est la recherche du ou des ordres totaux les plus "proches" d'un "tournoi".

Si le tableau C est un tableau de données $0 \leq c_{ij} \leq 1$ la maximisation de la fonction :

$$\sum_i \sum_j c'_{ij}\, y_{ij}$$

consiste à chercher la relation d'ordre total "approximant" au mieux une relation "floue" (les propriétés liées à la règle majoritaire ne sont plus systématiquement vérifiées).
Si par contre on pose $\hat{c}_{ij} = c_{ij}/m$ (où C est un tableau collectif construit à partir des opinions de m juges), le tableau collectif $\hat{C}$ sera bien du type $0 \leq \hat{c}_{ij} \leq 1$ et correspondra à la donnée de majorités en pourcentages; de ce fait la maximisation de $\hat{F}(Y) = \sum_i \sum_j \hat{c}'_{ij}\, y_{ij}$ sera équivalente à maximiser F(Y) sur le tableau C d'origine.

b) Cas où Y est une relation d'équivalence (partition)

i) Malgré la ressemblance entre les deux modèles que nous avons notée précédemment, la maximisation du nombre d'accords :

$$\sum_i \sum_j c_{ij}\, y_{ij}$$

n'est pas équivalente à la maximisation de la fonction F(Y). Dans ce cas ceci conduirait à l'obtention de la solution triviale suivante $y_{ij} = 1 \ \forall\ i$ et j, c'est à dire à la solution où tous les objets seraient situés dans la même classe.

Remarque:

On peut montrer, par contre, que la maximisation de la fonction F(Y) est équivalente à la maximisation de la fonction symétrisée de Y suivante:

$$\sum_i \sum_j (c'_{ij} + c'_{ji})\, y_{ij}$$

qui a l'avantage, de par la symétrisation du problème ainsi créée, d'éliminer du calcul la partie antisymétrique des données qui n' intervient pas dans la recherche d'une relation Y devant être par nature symétrique (voir [21] pour plus de détails).

ii) Si les tableaux individuels de similarités C^k représentent tous des partitions, ce problème particulier d'agrégation des similarités est connu sous le nom de : "Recherche d'une partition Centrale", ce problème a été posé par S.Regnier dans (24).

iii) Lorsque les tableaux individuels de similarités C^k ne représentent pas des partitions, même le cas d'un tableau C^k unique est intéressant.
Si ce tableau C^k est un tableau binaire quelconque, on se trouve dans le contexte d'un problème classique de théorie des graphes, consistant à approximer une relation binaire quelconque par une relation d'équivalence. Un exemple particulier d'un tel problème est le "Problème de Zahn", consistant à chercher la partition la plus proche d'une relation symétrique (25).
Si le tableau de données C prend ses valeurs dans (0-1) c'est à dire si $0 \leq c_{ij} \leq 1$, la maximisation de la fonction F(Y) associée

$$\sum_i \sum_j (2\, c_{ij} - 1)\, y_{ij}$$

consiste à chercher la relation d'équivalence "approximant au mieux" une relation de similarité "floue".
Ceci se rencontre souvent en classification automatique, lorsque l'on construit des tableaux carrés (nXn) de terme général s_{ij} i=1,2,...n j=1,2....n où les valeurs s_{ij} représentent la valeur d'indices de similarité calculés sur les n objets, ces indices variant de 0 à 1. En identifiant c_{ij} à s_{ij} on se retrouve dans le contexte précédent. Et

l'on maximise alors la fonction : $\sum_i \sum_j (2 s_{ij} - 1) y_{ij}$

(également dans ce cas, les propriétés liées à la règle de Condorcet ne sont pas toutes vérifiées, sauf pour certains types d'indices que l'on étudiera au Chapitre VII).

Dans ce cas également si l'on pose $\hat{c}_{ij} = c_{ij}/m$ (C étant un tableau collectif provenant de m juges), on a bien $0 \leq \hat{c}_{ij} \leq 1$ mais ceci revient à considérer les majorités en pourcentages et la maximisation de $\hat{F}(Y) = \sum \sum \hat{c}_{ij} y_{ij}$ est équivalente à la maximisation de la fonction $F(Y)$ construite sur le tableau collectif C d'origine.

V- LA METHODE DE RESOLUTION DU MODELE PROPOSE

5-1) QUELQUES METHODES EXISTANTES DE RESOLUTION EXACTE

a) Cas de l'agrégation des préférences

Dans le cas de l'agrégation d'ordres il existe un certain nombre de méthodes exactes; elles sont toutes fondées sur le principe d'énumération implicite des n! permutations (Branch and Bound). La plupart de ces méthodes sont détaillées ou citées dans (16). Malheureusement on peut considérer que ces méthodes ne sont applicables que pour des tailles de n (nombre d'objets) inférieures à 20.

b) Cas de l'agrégation de similarités

Pour ce problème il n'existait pas (à notre connaissance) de méthodes exactes de résolution jusqu'à aujourd'hui.

5-2) LA METHODE EXACTE PROPOSEE

Les deux modèles a) et b) du chapitreIV sont des modèles de programmation bivalente, c'est à dire en variables (0-1) et de ce fait sont a priori des problèmes excessivement compliqués vus leur tailles. Cependant nous avons pu obtenir des solutions optimales en des temps de calcul raisonnables, même pour des tailles de n relativement importantes.

Le principe de l'algorithme de résolution exacte que nous avons proposé repose sur l'utilisation de la programmation linéaire continue avec adjonction de coupes (cutting planes). Cet algorithme consiste alors à résoudre une succession de programmes linéaires continus associés aux modèles a) ou b), c'est à dire où l'on a négligé au préalable les contraintes d'intégrité $y_{ij} = 0$ ou 1, transformées en $0 \leq y_{ij} \leq 1$ et où l'on a ajouté à chaque étape les contraintes définissant les coupes successives. C'est un résultat bien connu de programmation linéaire

qu'une telle méthode est une méthode finie -chaque PL nécéssite un nombre fini d'itérations et le nombre de programmes linéaires est fini-.
Pourtant, à ce niveau deux questions se posent:

i) Combien de programmes linéaires faudra-t-il résoudre pour aboutir à une solution optimale bivalente ?
ii) La résolution d'un seul de ces programmes linéaires ainsi générés est-elle possible, compte tenu des tailles mises en oeuvre?

Contrairement à l'habitude pour ce type de problème, ce n'est pas le nombre de programmes linéaires successifs à résoudre qui sera la principale difficulté (puisqu'en fait le plus souvent on n'aura à traiter qu'un seul programme linéaire), mais bien plutôt la résolution d'un seul de ces programmes linéaires continus.

En effet la résolution des modèles a) et b) rendus continus en transformant la contrainte $y_{ij} = 0$ ou 1 par $0 \leq y_{ij} \leq 1$ est a priori impossible si on essaye de traiter directement ces programmes, car le programme sera automatiquement mis sous forme standard par adjonction de variables d'écart. Le nombre des contraintes étant alors de l'ordre de n^3, oblige a générer des matrices de bases de tailles $O(n^3) X O(n^3)$, ce qui bloque rapidement la place mémoire disponible, même pour n faible et rend impossible les calculs effectifs.

A ce niveau l'intérêt et l'originalité de la méthode que nous proposons est de pouvoir traiter le programme linéaire continu précédent avec des bases de tailles $O(n^2) X O(n^2)$ seulement. Dans (18) nous avons détaillé le processus qui nous a conduit à ce résultat. C'est la procédure "duale" du dual qui a été utilisée, avec le code produit-programme IBM MPSX/370 dont les avantages sont également détaillés dans (16).
Ceci nous a permis de résoudre de façon exacte des problèmes de tailles relativement grandes (jusqu'à n=80 pour l'instant).

Cependant, et ceci est un point fondamental, pour tous les exemples pratiques que nous avons traités (plus de 200 à l'heure actuelle, tant d'agrégation des préférences que d'agrégation des similarités), les solutions optimales des premiers programmes continus (c'est à dire sans adjonction de coupes) étaient toujours à valeurs (0-1), c'est à dire toujours solutions des modèles linéaires bivalents a) ou b). Le fait de n'avoir à résoudre la plupart du temps qu'un seul programme linéaire continu est un résultat empirique tout à fait remarquable, qui justifie amplement à lui tout seul l'approche "optimisation" que nous avons choisie.

En effet le nombre de programmes linéaires continus successifs que nous aurons à résoudre étant extrêmement

faible, sauf cas exceptionnels, l'algorithme exact que nous proposons permettra des temps de calculs très raisonnables. Ces temps de calcul sont détaillés au paragraphe suivant.

Remarque:

Pour les trop grandes tailles de n, nous avons la possibilité d'utiliser des heuristiques performantes (performance mesurée par rapport aux résultats de la méthode exacte sur les tailles de n ≤80), tant en agrégation des préférences qu'en agrégation des similarités (voir dans (16) la description de certaines de ces heuristiques).

5-3) TEMPS DE CALCUL

Nous donnons ici quelques temps de résolution (avec MPSX/370 sur 370/168 IBM incluant: Génération-Résolution-Edition)

TYPE du PB	TAILLE	NATURE	TEMPS(Meth.Exact.)	TEMPS(heuri.)
Archeologie	21	P	7 s	0.1 s
Sociologie	20	P	13 s	0.1 s
Medecine	23	P	15 s	0.2 s
Linguistique	25	C	7 s	0.2 s
Linguistique	28	C	11 s	0.2 s
Medecine	29	P	22 s	0.2 s
Linguistique	29	C	12 s	0.2 s
Esthetique	36	C	22 s	0.2 s
Psychologie	35	P	36 s	0.25 s
Sociologie	40	C	31 s	0.2 s
Test Compar.	40	P	297 s	0.35 s
Linguistique	45	C	60 s	0.25 s
Linguistique	45	C	58 s	0.25 s
Esthétique	48	C	65 s	0.30 s
Esthétique	54	C	117 s	0.30 s
Politique	54	P	220 s	0.90 s
Sociologie	72	C	157 s	0.90 s
Linguistique	77	C	331 s	1.20 s

(Dans la colonne NATURE "P" signifie que le résultat obtenu est une "partition", "C" que le résultat obtenu est un "classement").

VI- PROPRIETES GENERALES DES MODELES PRECEDENTS

Donnons tout d'abord une liste de propriétés intuitives qu'une méthode "cohérente" devrait pouvoir vérifier.Nous démontrerons ensuite que la méthode proposée précédemment

les possède toutes.

6-1) PROPRIETE DE CONSISTANCE ADDITIVE

Sans entrer dans une écriture mathématique fastidieuse de cette propriété, disons qu'elle correspond à l'idée suivante:
si à partir des m "juges" ou "critères" de départ on sélectionne un sous ensemble de m1 "critères" tels qu'on obtienne par application des méthodes précitées un résultat collectif "R" (soit un ordre O soit une partition P suivant le cas) et qu'à partir d'un sous ensemble disjoint du premier de m2 "critères" on obtienne toujours le même résultat collectif "R"; alors la propriété de CONSISTANCE ADDITIVE nous garantit qu'à partir de la réunion des 2 ensembles disjoints précédents de cardinal (m1+m2) on obtiendra toujours le même résultat collectif "R" (c'est à dire toujours l'ordre O ou la partition P).

6-2) PROPRIETE DE NEUTRALITE TOTALE

Là encore on peut définir cette propriété d'une façon simple: l'obtention du résultat cherché (qu'il soit un ordre collectif ou une partition collective) est indépendant simultanément de l'indiçage ou de la signification des "critères" ou "juges" servant à les classer ou les classifier, ainsi que de l'indiçage des objets. En d'autres termes, que le résultat doit rester inchangé quelle que soit la permutation des lignes ou des colonnes que l'on peut faire subir aux tableaux de données avant traitement. Cette propriété est souvent appelée: "Symmetric Neutrality" par les Anglo-Saxons; elle contient également comme cas particulier la propriété de "Non Labelled Structure" définie par S. Jackson (11).

6-3) PROPRIETE DE NON FIXATION DU NOMBRE DE CLASSES (pour les problèmes d'agrégation de similarités seulement)

Pour une méthode d'obtention d'une partition, vérifier cette propriété signifie simplement que la partition collective résultat sera obtenue sans fixation a priori du nombre de classes. Comme on le verra, la méthode proposée vérifie implicitement cette propriété.

6-4) VERIFICATION DE CES PROPRIETES

a) La propriété de "CONSISTANCE ADDITIVE" est vérifiée par la méthode proposée pour les deux types d'agrégation.

En effet, supposons que l'ensemble des "juges" se divise en deux sous-ensembles K1 et K2 tels que $K1 \cap K2 = \emptyset$ et

$|K1 \cup K2| = m$, puisque $c'_{ij} = \sum_{k \in K1} c'^{k}_{ij} + \sum_{k \in K2} c'^{k}_{ij}$

En posant $c'_{ij}(1) = \sum_{k \in K1} c'^{k}_{ij}$ et $c'_{ij}(2) = \sum_{k \in K2} c'^{k}_{ij}$, et la représentation des contraintes sur Y par $Y \in R$; si Y^* est solution de

$\underset{Y \in R}{\text{Max}} \sum_{i,j} c'_{ij}(1)\, y_{ij}$ et simultanément de $\underset{Y \in R}{\text{Max}} \sum_{i,j} c'_{ij}(2)\, y_{ij}$

alors Y^* sera également solution optimale de $\underset{Y \in R}{\text{Max}} \sum c'_{ij}\, y_{ij}$

ceci étant du au fait que $c'_{ij} = c'_{ij}(1) + c'_{ij}(2)$ et à l'inégalité suivante:

$$\underset{Y \in R}{\text{Max}} \sum c'_{ij} y_{ij} \leq \underset{Y \in R}{\text{Max}} \sum c'_{ij}(1) y_{ij} + \underset{Y \in R}{\text{Max}} \sum c'_{ij}(2) y_{ij}$$

b) La propriété de NEUTRALITE TOTALE est vérifiée par la méthode proposée pour les deux types d'agrégation

En effet par définition d'un programme linéaire où les indices i et j jouent le rôle d'indices muets et où l'indice k est implicitement contenu dans les données c' , la propriété de NEUTRALITE TOTALE est vérifiée également.

c) La propriété de NON FIXATION du nombre de classes est vérifiée par la méthode dans le cas de l'agrégation de similarités.Ceci par construction.

Remarque:

La question qui vient à l'esprit ici est la suivante: Certaines des méthodes actuellement utilisées en classification automatique pourraient-elles ne pas vérifier les propriétés précédentes?
la réponse est: oui.

En effet considérons le problème de classification hiérarchique associé au tableau de données binaires suivant:

	V1	V2	V3	V4	V5	V6	V7	V8
a	1	0	1	0	0	0	0	0
b	1	1	0	0	1	0	1	1
c	1	1	0	1	1	0	0	1
d	0	1	1	1	1	1	0	0

Supposons que nous divisions l'ensemble des variables en deux sous-ensembles disjoints soit: K1=(V1,V2,V3,V4) et K2=(V5,V6,V7,V8)

On obtient alors deux sous tableaux de données, pour

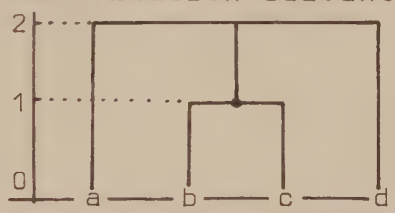
lesquels on peut calculer deux tableaux de distances mutuelles entre les points (a,b,c,d) (calculés à partir de la distance d1 entre les profils)

On obtient les deux tableaux de distances suivants:

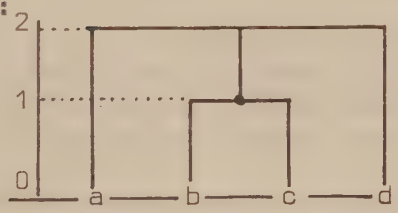
	a	b	c	d
a		2	3	3
b	2		1	3
c	3	1		2
d	3	3	2	

	a	b	c	d
a		3	2	2
b	3		1	3
c	2	1		2
d	2	3	2	

A partir du premier tableau de distances construit sur l'ensemble K1 on obtient, en utilisant l'algorithme de Johnson ou tout autre algorithme donnant l'ultramétrique sous-dominante, le résultat présenté sous la forme de l'arbre de classification suivant:



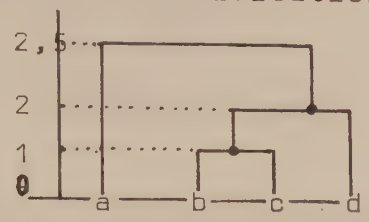
A partir du deuxième tableau de distances construit sur l'ensemble K2 on obtient le même arbre de classification c'est à dire:



Puisqu'à partir des sous-ensembles de variables K1 et K2 on obtient 2 fois le même arbre de classification, en toute logique la classification des points (a,b,c,d) par rapport à la réunion K1UK2 devrait nous redonner le même arbre de classification.
Or ce n'est précisemment pas le cas; en effet, le tableau de distances mutuelles entre les 4 points, construit à partir des données sur les 8 variables est le suivant:

	a	b	c	d
a		5	5	5
b	5		2	6
c	5	2		4
d	5	6	4	

L'algorithme de Johnson appliqué à ce tableau de distances nous donne l'arbre de classification suivant:



(les valeurs de niveaux sont données ici après division par 2 des valeurs du tableau de distances)

Cet arbre de classification est différent du précédent car "a" et "d" ne se réunissent plus au même niveau mais à deux niveaux différents.

La conclusion de ce petit exemple est que certaines méthodes de classification hiérarchique ne vérifient pas la propriété de CONSISTANCE ADDITIVE.

De la même façon, la propriété de NEUTRALITE TOTALE n'est pas vérifiée par certains algorithmes de classification non hiérarchique. (ISODATA ou certaines versions de l'algorithme des "nuées dynamiques" pour lesquelles on tire au hasard les "patrons").

VII- PRISE EN COMPTE DES DONNEES

La présentation sous forme de tableaux de comparaisons par paires C^k permet une grande souplesse dans la prise en compte des données sans **transformations a priori des données brutes, par le biais de l'approche bi-dimensionnelle du codage.**

Dans la mesure où nous avons présenté dans d'autres articles (voir (16) et (19) par exemple) comment transformer en données relationnelles des données de préférences, nous allons insister presque exclusivement dans cet article sur les données de similarités.

En fait les données sur lesquelles nous allons travailler se divisent en deux types distincts:

- les données sont déjà directement sous forme de tableaux de comparaisons par paires;

- elles ne sont pas initialement sous forme de tableaux de comparaisons par paires et il s'agit alors de les mettre implicitement ou explicitement sous forme de tableaux C^k

(binaires) ou de tableau C.

7-1) DONNEES DIRECTEMENT SOUS FORME DE TABLEAUX DE COMPARAISONS

a) Cas de l'agrégation des préférences.

Se présentent directement sous forme de tableaux de comparaisons par paires les types de données suivants:

1- les tableaux de tournois (football, escrime, rugby,...);
2- les graphes de "dominations" ou de hiérarchies dans les populations animales ou humaines (voir exemple sur les chimpanzés dans (16));
3- les graphes de relations hiérarchiques dans les entreprises (la distribution des 1 et des 0 dans de tels tableaux peut être quelconque);
4- les graphes représentation de "systèmes";
5- les données de préférences par paires (enquêtes psychosociologiques , enquêtes d'opinions, etc...).

Remarque:

Cette présentation directe des données par comparaisons par paires permet de prendre en compte les données paradoxales suivantes:

i) EFFET CONDORCET INDIVIDUEL
Pour le "juge" k on peut très bien avoir la situation suivante: le juge k préfère x à y, y à z et z à x, c'est à dire avoir:

$c^k_{xy} = 1$, $c^k_{yz} = 1$, $c^k_{zx} = 1$

ii)EFFET DE BORD
Un juge k ne donne pas un "classement" de n objets avec précision (si n grand), mais il connait avec précision ses "meilleurs" et ses "plus mauvais", exemple:

X----Y----Z .....ne sait pas ..... X'----Y'----Z'
1er 2ème 3ème (n-2) (n-1) nième

b) Cas de l'agrégation de similarités

Se présentent directement sous forme de tableaux de comparaisons par paires.

1-Les tableaux individuels de similarités
Chaque "juge" définit directement ses "similarités" sur une population de n objets en posant:

soit : $c^k_{ij} = 1$ si pour lui i et j sont "similaires" ou "semblables";

$c^k_{ij} = 0$ sinon.

ou avec plus de nuance:

$c^k_{ij} = 1$ si pour lui les objets i et j sont "similaires";

$c^k_{ij} = 1/2$ s'il ne "sait pas";

$c^k_{ij} = 0$ dans les autres cas.

(Voir dans (19) un exemple de tels tableaux de similarités relatifs à l'opinion de 51 personnes sur les ressemblances entre 12 machines à écrire)

Remarque:

En général ces tableaux sont symétriques mais rarement transitifs, c'est à dire ne représentent pas forcément une partition. Ceci correspond à la prise en compte de l'EFFET POINCARE individuel: où un "juge" k trouve que X et Y sont similaires, Y et Z similaires mais pas forcément X et Z.

2-Les tableaux carrés (nXn) de distances, d'éloignements, de similarités, ou de noircissement du type BERTIN-CZEKANOWSKY

Ce sont en général des tableaux résultats de calculs préalables pour lesquels nous ne connaissons pas "a priori" le tableau de données brutes, croisant les n objets avec eux-mêmes. Le terme (i,j) de ce tableau soit s_{ij} représente:
-soit une mesure de distance ou d'éloignement entre i et j (non nécessairement symétrique);
-soit une mesure de similarité entre i et j;
-soit une valeur de "grisé" caractérisation visuelle d'une liaison entre i et j.

i)Cas d'un tableau de distance ou d'éloignement
On peut définir un tableau de similarité individuel C^k en posant:

$c^k_{ij} = 1$ Si $s_{ij} \leq s$ (s étant un seuil donné);

$c^k_{ij} = 0$ Sinon.

C^k peut très bien ne pas représenter une partition ou n'être pas symétrique si S ne l'est pas lui-même.

ii) Soit S un tableau de similarités (construit à partir d'un indice) tel que $0 \leq s_{ij} \leq 1$, alors on fabriquera

directement le tableau C' en posant:

$$c'_{ij} = 2\, s_{ij} - 1$$

et l'on maximisera la fonction économique suivante:

$$\Sigma\ \Sigma (2s_{ij} - 1)\ y_{ij}$$

mais cette fonction n'est pas le résultat de l'application d'une règle de majorité collective et ne verifiera pas sauf exception les propriétés déjà présentées auparavant.

iii) Dans le cas des tableaux de BERTIN-CZEKANOWSKY, c'est le noircissement des cases qui fait office de tableaux de similarités, on posera par exemple pour un tableau C' :

c'_{ij} = 1 Si la case (i,j) est noire;

c'_{ij} = 0 Si elle est grise;

c'_{ij} = -1 Si elle est blanche.

Plus généralement s'il y a k niveaux de gris on construira directement le tableau collectif C en posant:

c_{ij} = k Si la case (i,j) est totalement noire;

c_{ij} = k-1 Si la case est gris foncé;

....

cij = 0 Si elle est complétement blanche.

L'idée sous-jacente ici étant que la valeur d'un grisé est additive et correspond à l'addition de plusieurs critères (virtuels) jouant sur du (noir et blanc). Ainsi un gris de valeur (k-2) par exemple pour la case (i,j) correspond au fait que k-2 "critères" ont jugé (i,j) noirs (c'est à dire ressemblants) alors que deux "critères" (virtuels) les ont trouvés blancs c'est à dire "non similaires".

7-2) DONNEES NON DIRECTEMENT SOUS FORME "COMPARAISONS PAR PAIRES"

Dans ce cas il s'agit de transformer les données de façon à obtenir des tableaux C^k (représentant un "juge" ou un "critère") sous formme de comparaisons par paires afin de pouvoir utiliser les algorithmes précédemment décrits.

a) Cas de l'agrégation des préférences

Nous ne donnerons ici, à titre d'exemple que la façon dont on transforme en tableaux de comparaisons par paires des données initialement sous forme de "classements de n objets" (ordres totaux).

Supposons que m "juges" ou "critères" classent n objets . Alors en notant r_{ik}= rang de l'objet i pour le juge k, on obtient le tableau C^k associé de la façon suivante:

$c^k_{ij} = 1$ Si $r_{ik} \leq r_{jk}$;
$c^k_{ij} = 0$ Sinon.

Dans ce cas particulier, si l'on permute les lignes et les colonnes du tableau C^k obtenu en fonction du classement du juge k on obtient des "1" au dessus de la diagonale de C^k et des "0" en dessous

De la même façon moyennant quelques variantes on peut obtenir un tableau C^k associé à des données initiales sous forme de préordres totaux, de quasi-ordres, d'ordres partiels, d'échelles de notes etc.... (Voir dans (20), le détails de ces transformations et dans (16) une discussion sur l'absence de perte d'information et les avantages de telles transformations sur les données).

b) Cas de l'agrégation de similarités

1) Les données initiales sont des partitions

Le cas le plus simple correspond à la situation où un "juge" ou un "critère" représente une variable à plusieurs "modalités". Le tableau de données brutes associé est alors un tableau de modalités, où le terme général r_{ik} représente la valeur de la modalité attribuée à l'objet i par le juge ou critère k. Le tableau C^k est alors défini de la façon suivante:

$c^k_{ij} = 1$ Si $r_{ik} = r_{jk}$ c'est à dire si i et j ont même modalité pour le juge k;

$c^k_{ij} = 0$ Sinon.

C^k représente alors une relation d'équivalence (partition).

Un autre codage possible, plus souple, utilisant directement le tableau C'^k permet à ce niveau de prendre en compte des variables pour lesquelles il existe une hiérarchie sous-jacente dans les modalités

$c'^k_{ij} = 1$ Si $|r_{ik} - r_{jk}| \leq s$;
$c'^k_{ij} = 0$ Si $|r_{ik} - r_{jk}| = s$;
$c'^k_{ij} = -1$ Si $|r_{ik} - r_{jk}| > s$.

s est un seuil fixé à l'avance en général faible (s=1 ou 2).

Le tableau C'^k dans ce cas ne représente pas une partition, puisqu'il contient par définition des configurations où l'effet Poincaré se produit.

Ce type de données (variables à modalités) est extrèmement fréquent en analyse des données lorsqu'il y a hétérogénéité des variables. En effet prenons l'exemple suivant:
Si l'on veut classifier par exemple une population d'individus pour lesquels on a des informations sur le SEXE, le REVENU, la catégorie SOCIO-PROFESSIONNELLE; le problème se pose d'unifier les variables. En effet le SEXE représente une partition des individus en deux classes, la catégorie SOCIO-PROFESSIONNELLE une partition en 3 ou 4 classes. Mais le REVENU est une variables continue qu'ilconvient de transformer en variables à modalités pour la circonstance. De ce fait c'est le type de variable le plus discontinu qui l'emporte et qui oblige à transformer les autres variables en données de "partitions". Cette configuration de problème pathologique ou délicate pour les autres méthodes d'analyse des données est le cas le plus adéquat et le plus simple dans le cadre de l'approche proposée ici.

2) Cas où les données initiales sont des données binaires de (présence - absence) ou de réponses (oui-non)

Dans le cas où le "juge" ou "critère" k représente une propriété ou une question, le tableau de données brutes associé est alors un tableau de (présence-absence) ou de réponses (oui-non) où:
r_{ik}= 1 Si l'objet i a la propriété k ou a répondu "oui" à la question k;
r_{ik}= 0 Sinon.

La notion de similarité peut se définir de plusieurs façons dépendant du contexte du problème.

i) 1er Cas

c'^k_{ij} = 1 Si $r_{ik} - r_{jk} = 0$ c'est à dire que i et j ont ou n'ont pas simultanément la propriété k

c'^k_{ij} = -1 Si $r_{ik} - r_{jk} \neq 0$

ii) 2ème Cas

c'^k_{ij} = 1 Si $r_{ik}.r_{jk} = 1$ c'est à dire que i et j ont tous les deux la propriété k

c'^k_{ij} = -1 Sinon c'est à dire si $r_{ik}.r_{jk}=0$

iii) 3ème Cas

$c'^{k}_{ij} = 1$ Si $r_{ik} = r_{jk} = 1$, i et j ont simultanément la propriété k,

$c'^{k}_{ij} = 0$ Si $r_{ik} = r_{jk} = 0$, i et j n'ont pas simultanément la propriété k,

$c'^{k}_{ij} = -1$ Si $r_{ik} - r_{jk} \neq 0$.

Cette possibilité de définir le tableau C'^{k} de plusieurs façons différentes correspond au poids ou à la signification que l'on doit accorder suivant le contexte aux configurations du type (0-0). Point dont la multitude d'indices de similarités définis par les spécialistes de la classification montre le caractère ambigu. Nous aurons d'ailleurs l'occasion de revenir plus loin sur la relation entre indices de similarités, tableaux C'^{k} et règle de majorité.

Interprétation de ces différentes configurations

1) Le 1er cas correspond à la configuration où chaque "propriété k" induit une partition en 2 classes seulement.
Exemple:
la propriété "SEXE" (1-1) implique que i et j sont 2 hommes
(0-0) implique que i et j sont 2 femmes

donc dans les 2 cas i et j seront dans la même classe.

2) Le 2ème cas correspond à la configuration où pour deux individus i et j ne pas avoir une propriété k équivaut à une dissimilarité entre eux. La propriété k n'induit pas une partition en 2 classes.

3) Le 3ème cas correspond à la configuration où le fait que i et j n'aient pas la propriété k n'implique pas forcément que i et j se ressemblent mais n'implique pas plus leur non ressemblance (c'est à dire implique le doute).
Exemple:
la propriété "AVOIR LES YEUX VERTS"
Dans ce cas les configurations (1-1) ==> i et j ont les yeux verts
(0-1) ==> i a les yeux verts, j n'a pas les yeux verts

sont des configurations sans ambiguité.

Mais la configuration (0-0) ==> i et j n'ont pas les yeux verts. Mais i et j peuvent très bien avoir tous les deux les yeux bleus, donc se ressembler.

est ambigüe.
Dès lors poser $c'^{k}_{ij} = 0$ si $r_{ik} = r_{jk} = 0$, traduit cette

ambiguité. Puisque $c'^{k}_{ij} = 0$ cette valeur n'entrera pas en ligne de compte dans les calculs, contrairement aux deux cas précédents.

3) Le tableau de données brutes est un tableau quantitatif

Ce tableau correspond à des données représentant des "mesures" sur des variables continues mais pas forcément du même type. Dans ce cas, comme dans d'autres méthodes, une étude statistique descriptive préalable est nécessaire (fabrication d'histogrammes). Ces histogrammes associés à chaque "variable" ou "critère" k induisant alors des classes sur la population. On peut alors comme précédemment définir la similarité par les tableaux C^k ou C'^k suivants:

$c^k_{ij} = 1$ Si i et j sont dans une même classe de l'histogramme associé à la variable k

$c^k_{ij} = 0$ Sinon.

Dans ce cas C^k représente une partition.

Autre codage possible moins discontinu:

$c'^k_{ij} = 1$ Si i et j sont dans des classes adjacentes de l'histogramme associé à k, c'est à dire si $|\text{Num classe de i} - \text{Num classe de j}| < s$

$c'^k_{ij} = 0$ Si $|\text{Num classe de i} - \text{Num classe de j}| = s$

$c'^k_{ij} = -1$ Si $|\text{Num classe de i} - \text{Num classe de j}| > s$

(s étant un seuil faible, fixé a priori).
Dans ce cas C'^k n'est pas un tableau représentant une partition.

Exemple : Mesures sur des squelettes en zoologie (craniométrie par exemple).

Comme nous venons de le voir, l'utilisation de tableaux de données sous forme de comparaisons par paires permet donc de prendre en compte des données aussi différentes que les tournois sportifs ou les résultats de tests psychométriques (données 0 - 1).
Les tableaux C^k ou C'^k eux mêmes peuvent tout aussi bien représenter des données aussi simples que des ordres totaux ou des partitions, que des données complexes ou paradoxales (Effet Condorcet ou Poincaré, impossible à prendre en compte autrement).
Outre cette souplesse d'utilisation, la prise en compte des données sous forme de tableaux C^k ou C'^k permet d'eviter des codages "a priori" de données souvent arbitraires et sans

signification.
Nous allons illustrer cette affirmation par la mise en évidence des relations existant entre les différents indices de similarité utilisés habituellement dans le cas de données binaires et les tableaux C^k, C'^k, C ou C'.

7-3) RELATION ENTRE LES TABLEAUX C , C' , LA REGLE MAJORITAIRE DE CONDORCET ET DIFFERENTS INDICES DE SIMILARITE (COEFFICIENTS D'ASSOCIATION)

Supposons que l'on ait extrait d'un tableau nXm de données binaires de présence - absence les lignes i et j (correspondant à deux objets quelconques relativement à m critères), on a alors sa configuration suivante:

	V1	V2	V3	V4	V5	V6	....	Vm
i	0	1	1	0	1	0	1	1
j	1	0	1	0	0	0	1	0

On peut former alors le tableau 2X2 suivant:

		j = 1	j = 0
i	1	11(i,j)	10(i,j)
	0	01(i,j)	00(i,j)

où:
11(i,j) = nombre de critères ou propriétés que i et j possèdent simultanément,
01(i,j) = nombre de propriétés que i ne possède pas mais que j possède, etc...
Dans le cas présent,
11(i,j) = 2, 10(i,j) = 3, 01(i,j) = 1, 00(i,j) = 2
Les différents coéfficients d'association ou indices de similarité connus sont des fonctions de 11(i,j), 10(i,j), 01(i,j), 00(i,j).

a) Présentation de différents indices de similarité:

1) Premier TYPE

Les indices qui suivent varient entre 0 et 1 (1 si i et j sont semblables, 0 si i totalement différent de j). Ils font jouer dans la valeur de l'indice un rôle "positif" équivalent à (1-1) et (0-0) et "négatif" à (0-1) et (1-0).

i) indice de SOKAL et MICHENER (1958):

$$Ssm(i,j) = \frac{11(i,j) + 00(i,j)}{m} = \frac{\text{nombre de concordances}}{\text{nombre de variables}}$$

ii) indice de SOKAL et SNEATH (1968):

$$Sss(i,j) = \frac{2\,(11(i,j) + 00(i,j))}{11(i,j) + 00(i,j) + m}$$

iii) indice de ROGERS et TANIMOTO (1960)

$$Srt(i,j) = \frac{11(i,j) + 00(i,j)}{2m - (11(i,j) + 00(i,j))}$$

2) 2ème TYPE

Les indices qui suivent varient entre 0 et 1 . Ils font jouer un rôle "positif" à (1-1) et font jouer à la configuration (0-0) un rôle "négatif" identique à celui des configurations (1-0) et (0-1). (On a rajouté pour l'occasion 2 indices virtuels qui n'existent pas dans la littérature par soucis d'homogénéité avec le reste des indices utilisés).

i) Indice de RAO1

$$Sr1(i,j) = \frac{11(i,j)}{m} = \frac{\text{Nombre de concordances (1-1)}}{\text{Nombre de variables}}$$

De la même façon que pour les indices du premier type on peut construire les deux indices suivants:

ii) Indice de RAO2 (°)

$$Sr2(i,j) = \frac{2x(11(i,j))}{11(i,j) + m}$$

iii) Indice de RAO3 (°)

$$Sr3(i,j) = \frac{11(i,j)}{2m - 11(i,j)}$$

3) 3ème TYPE

Les indices qui suivent varient également de 0 à 1. Nous les avons introduits parce qu'ils jouent un rôle intermédiaire entre les indices du 1er type et ceux du 2ème type. Pour

ces indices la configuration (1-1) joue un rôle "positif", la configuration (0-0) un rôle "neutre" et (0-1) - (1-0) un rôle "négatif".

i) Indice de MARCOTORCHINO-MICHAUD (MM1)

$$Smm1(i,j) = \frac{11(i,j) + 1/2(00(i,j))}{m}$$

On a la relation:

$$Smm1(i,j) = 1/2 \ (\ Ssm(i,j) + Sr1(i,j) \ )$$

ii) Indice MM2

$$Smm2(i,j) = \frac{2(\ 11(i,j) + 1/2(\ 00(i,j)) \ )}{11(i,j) + 1/2 \ (00(i,j)) + m}$$

iii) Indice MM3

$$Smm3(i,j) = \frac{11(i,j) + 1/2 \ (00(i,j))}{2m - (11(i,j) + 1/2(00(i,j)))}$$

4) 4ème TYPE

Les 4 indices qui suivent (3 existants + 1 virtuel) sont très souvent utilisés dans la littérature; ils varient également de 0 à 1. Ils font jouer un rôle "positif" à la configuration (1-1), un rôle "négatif" aux configurations (1-0) et (0-1) et ne font jouer "aucun" rôle à la configuration (0-0). Comme on le verra ultérieurement, ils sont de plus étroitement liés à la mise sous forme "disjonctive complète" des tableaux de données à modalités. De ce fait nous les présentons ici dans un ordre non intuitif a priori, mais dont l'interprétation montrera la nécessité et la logique.

i) Indice de DICE

$$Sd(i,j) = \frac{11(i,j)}{11(i,j) + 1/2 \ (10(i,j) + 01(i,j))}$$

ii) Indice de DICE bis (°)

$$Sdb(i,j) = \frac{11(i,j)}{11(i,j) + 1/4 \ (10(i,j) + 01(i,j))}$$

iii) Indice de JACCARD

$$Sj(i,j) = \frac{11(i,j)}{11(i,j) + 10(i,j) + 01(i,j)}$$

iiii) Indice de SORENSEN

$$Sso(i,j) = \frac{11(i,j)}{11(i,j) + 2\ (10(i,j) + 01(i,j))}$$

Ces différents indices sont fonction les uns des autres suivant le même schéma homographique dans chacun des 4 types. Comme le montre le tableau suivant:

TYPE 1	TYPE 2
$Ssm = Ssm$	$Sr1 = Sr1$
$Sss = \frac{2\ Ssm}{Ssm + 1}$	$Sr2 = \frac{2\ Sr1}{Sr1 + 1}$
$Srt = \frac{Ssm}{2 - Ssm}$	$Sr3 = \frac{Sr1}{2 - Sr1}$
TYPE 3	**TYPE 4**
$Smm1 = Smm1$	$Sd = Sd$
$Smm2 = \frac{2\ Smm1}{Smm1 + 1}$	$Sdb = \frac{2\ Sd}{Sd + 1}$
$Smm3 = \frac{Smm1}{2 - Smm1}$	$Sj = \frac{Sd}{2 - Sd}$

On a de plus la relation suivante entre les indices Smm1, Ssm et Sr1:

$$Smm1 = 1/2\ (\ Ssm + Sr1\ )$$

b) Relations entre ces différents indices de similarités et les différents codages des tableaux de (présence-absence).

1) Relation entre Ssm, Sss et Srt et le "1er mode" de codage
(voir 7-2 -b-2)

Dans ce cas $c_{ij} = \sum_k c^k_{ij}$ = Nombre de 1 et de 0 communs à i et j

$= 11(i,j) + 00(i,j)$

on a donc:

$$Ssm = \frac{c_{ij}}{m} = \frac{c'_{ij} + m}{2m}$$

Propriété (1-1)
L'indice de SOKAL et MICHENER n'est rien d'autre que la "fréquence" du nombre de "critères" ou "juges" ayant considéré que i et j étaient "semblables".

Reciproque (1-1)
Si l'on connait le nombre de "critères" d'un problème et le tableau de similarités associé, construit à partir de l'indice de Sokal et Michener on a directement le tableau collectif C (sans avoir besoin des données brutes) en posant:

$$c_{ij} = m\ Ssm\ (i,j)$$

Propriété (1-1) bis
Il est dès lors évident que $c'_{ij} \geq 0$ (la règle de la majorité par paire est vérifiée $\Longleftrightarrow c_{ij} \geq m/2$ et $Ssm(i,j) \geq 1/2$

donc: $Ssm \geq 1/2 \Longleftrightarrow c'_{ij} \geq 0$

Propriété (1-2)
De la même façon compte tenu de ce qui précéde, l'indice de SOKAL et SNEATH Sss(i,j) est une fonction homographique du terme général c_{ij} du tableau C:

$$Sss = \frac{2\ c_{ij}}{c_{ij} + m}$$

Réciproque (1-2)
Si l'on connait le nombre de "critères" ou de "juges" d'un problème et un tableau de similarités construit à partir de l'indice de SOKAL et SNEATH, on a directement le tableau collectif C en posant:

$$c_{ij} = \frac{m\ Sss\ (i,j)}{2 - Sss\ (i,j)}$$

Propriété (1-2) bis
On voit d'après la formule précédente que si la règle de la majorité de Condorcet est vérifiée par paire, c'est à dire si $c'_{ij} \geq 0 \Longleftrightarrow c_{ij} \geq m/2$ on a:

$$\frac{Sss}{2 - Sss} \geq 1/2 \quad \Longrightarrow \quad Sss(i,j) \geq \frac{2}{3}$$

Il faut donc que l'indice de SOKAL et SNEATH soit supérieur à 2/3 pour que la règle majoritaire sur une paire soit vérifiée.

Réciproque (1-2) bis
Réciproquement dire que $Sss(i,j) \geq 1/2$ implique :

$$\frac{2\, c_{ij}}{c_{ij} + m} \geq \frac{1}{2} \quad \Longrightarrow \quad c_{ij} \geq \frac{m}{3}$$

Considérer que 2 objets sont "semblables" si l'indice de SOKAL et SNEATH est supérieur à 1/2, revient à appliquer une règle de majorité "au tiers".

L'indice de SOKAL et SNEATH est donc plus "généreux" que l'indice de SOKAL et MICHENER.

Propriété (1-3)
De la même façon que pour les 2 indices précédents, l'indice de ROGERS et TANIMOTO est une fonction homographique du terme général cij du tableau collectif C, construit à partir du 1er mode de codage. On a alors:

$$Srt = \frac{c_{ij}}{2m - c_{ij}}$$

Réciproque (1-3)
Si l'on possède, comme données, le nombre de "critères" en jeux et un tableau de similarités construit à partir de l'indice de ROGERS et TANIMOTO, on a directement le tableau collectif C en posant:

$$c_{ij} = m \frac{2\, Srt(i,j)}{Srt(i,j) + 1}$$

Propriété (1-3) bis
On voit d'après la formule précédente que si la règle de majorité simple par paire est vérifiée, c'est à dire si $c_{ij} \geq m/2$, on a:

$$\frac{2\, Srt}{Srt + 1} \geq \frac{1}{2} \quad \Longrightarrow \quad Srt \geq \frac{1}{3}$$

Réciproque (1-3)bis
Réciproquement dire que $Srt \geq 1/2$ implique:

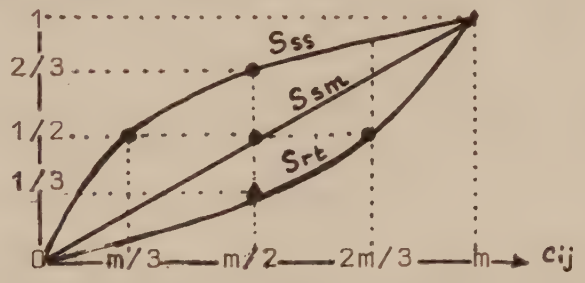
$$\frac{c_{ij}}{2m - c_{ij}} \geq \frac{1}{2} \quad ===) \quad c_{ij} \geq \frac{2\,m}{3}$$

Donc considérer que 2 objets i et j sont "similaires" si l'indice de ROGERS et TANIMOTO est supérieur à 1/2 revient à appliquer une règle de majorité par paire "aux deux-tiers".

L'indice de ROGERS et TANIMOTO est donc un indice plus "sévère" que l' indice de SOKAL et MICHENER.

En conclusion , ces différents indices proposés dans la littérature ne sont que des variations sur le thème de la règle de la majorité. La règle de la majorité simple ayant de nombreuses "bonnes" propriétés (en particulier la vérification de la Consistance additive), ceci explique le côté très "empirique" et très "approximatif" des processus de fabrication d'indices jouant sur des règles non évidentes et intuitives (règle au "tiers" ou au "deux-tiers").

Remarque:
On a de plus le graphique résumé suivant:



L'écart entre Sss et Srt soit (Sss - Srt) est une fonction de c_{ij} donnée par:

$$\frac{3\,c_{ij}\,(m - c_{ij})}{2\,m^2 + c_{ij}\,(m - c_{ij})}$$

cet écart est maximum pour $c_{ij} = m/2$, et il vaut 1/3.
Or comme le voisinage de la majorité simple est une zône critique où justement les 3 indices Ssm, Sss et Srt sont les plus différents, ceci confirme bien le côté arbitraire de certaines transformations a priori sur les données.

2) Relation entre Sr1 , Sr2 , et Sr3 et les tableaux C construits à partir du "2ème mode" de codage.

Dans ce cas on a: $c_{ij} = \sum_k c_{ij}^k = 11(i,j)$

Pour ce codage, les indices de RAO1, RAO2 et RAO3 jouent les rôles tenus respectivement par les indices de SOKAL-MICHENER, SOKAL-SNEATH, ROGERS-TANIMOTO pour le codage précédent. Les Propriétés (1-1) et (1-1) bis, (1-2) et (1-2) bis, (1-3) et (1-3) bis et leurs réciproques se transposent automatiquement dans ce cas en des Propriétés notées (2-1), (2-2) (2-3) etc.. pour lesquelles il suffit de changer le noms des indices et la signification du codage du tableau C, sans en changer l'écriture.

3) Relation entre les indices de similarités Smm1, Smm2 et Smm3 et le "3ème mode" de codage des tableaux C^k et C'^k.

Rappelons que ce codage se ramène à:

$c^k_{ij} = 1$ ou $c'^k_{ij} = 1$ pour la configuration (1-1)

$c^k_{ij} = 1/2$ ou $c'^k_{ij} = 0$ pour la configuration (0-0)

(ceci implique que le terme c'^k_{ij} correspondant ne sera pas pris en compte dans la fonction objectif F(Y))

$c^k_{ij} = 0$ ou $c'^k_{ij} = -1$ pour les configurations (0-1) et (1-0)

Dans ce cas:

$$c_{ij} = \sum_k c^k_{ij} = \sum_{k|(1-1)} c^k_{ij} + \sum_{k|(0-0)} c^k_{ij}$$

$$c_{ij} = 11(i,j) + 1/2\ (00(i,j))$$

Pour ce codage, les indices de similarités de MARCOTORCHINO-MICHAUD1, MM2 et MM3, jouent le rôle respectif des indices de SOKAL-MICHENER, SOKAL-SNEATH, ROGERS-TANIMOTO pour le "premier mode" de codage et de ce fait les Propriétés (1-1),(1-2),(1-3) et leurs réciproques se transposent directement en des Propriétés notées (3-1),(3-2),(3-3) etc.. , en changeant seulement le nom des indices et en utilisant le nouveau codage de C^k.

4) Relation entre les Indices de DICE - JACCARD - SORENSEN calculés sur un codage sous forme disjonctive complète de tableaux de données tel qu'il est préconisé par J.P. BENZECRI (5) Tome II pp. 23 et 24 en cas de variables hétérogènes et le codage des tableaux Ck associés à des variables à modalités tel que nous l'avons présenté en 7-2-b-1.

Considérons par exemple 4 variables, la première ayant 2 modalités (notée 1-2), la deuxième 3 modalités, la troisième 3 modalités, la quatrième 4 modalités, et les modalités

notées 1 ou 2 ou 3 ou 4 de deux individus i et j parmi n, associés à ces variables.
On a alors le tableau de données direct suivant: (nX4)

	V1	V2	V3	V4
i	1	3	2	4
j	2	3	3	3

Le codage sous forme disjonctive complète de ce tableau voir (5) p. 24 conduit au nouveau tableau de données (nX12) suivant:

	V11	V12	V21	V22	V23	V31	V32	V33	V41	V42	V43	V44
i	1	0	0	0	1	0	1	0	0	0	0	1
..	.	.	.	.					.	.	.	.
j	0	1	0	0	1	0	0	1	0	0	1	0

D'après le premier tableau de données (codage partition) on déduit le tableau collectif de comparaisons par paires C en posant:

c_{ij} = Nombre de variables pour lesquelles i et j ont la même modalité.

De la même façon, à partir du deuxième tableau sous forme disjonctive complète, on déduit le résultat suivant:
Nombre de cas où i et j ont la même modalité = 11(i,j)
Il est alors évident que:
c_{ij} (codage partitions) = 11(i,j) (codage forme disjonctive complète)

De la même manière, puisqu'il y a automatiquement un "1" et un seul par profil de l'individu pour chaque variable k.
On a le résultat suivant:
1/2(01(ij) + 10(i,j)) = Nombre de variables pour lesquelles (codage disjonctif) i et j n'ont pas les mêmes modalités.

D'où:

$$Sd\ (i,j) = \frac{11(i,j)}{11(i,j)+1/2(01(i,j)+10(i,j))} = \frac{c_{ij}}{c_{ij}+(m-c_{ij})}$$

ici $Sd(i,j) = 1/4 = \frac{c_{ij}}{m}$

Propriété (4-1)
L'indice de DICE calculé sur un tableau de données mises sous forme disjonctive complète n'est rien d'autre que la fréquence du nombre de variables ayant la même valeur des modalités pour i et j dans le "codage partition" proposé précédemment. Utiliser l'indice de DICE revient à appliquer la règle de la majorité lorsque l'on conserve ces tableaux sous forme "codage partition"; et l'on a:

c_{ij} = m (Nombre de variables)x Sd (i,j)

Donc appliquer la règle de la majorité sur le tableau de données brutes (partitions) revient à considérer que Sd $\geq 1/2$ sur le même tableau mis sous forme disjonctive complète.

Remarque: Il est à noter qu'ici le tableau C collectif joue le rôle de ce qu'on appelle tableau de BURT dans la terminologie de J.P.Benzecri, mais un tableau de BURT portant sur une correspondance K(I,I) où le terme général k(i,j) = nombre de critères pour lesquels i et j ont la même modalité c'est à dire c_{ij} dans notre présentation.

La forme disjonctive complète n'apporte strictement rien vis à vis de la méthode proposée mais elle est nécessaire en Analyse Factorielle des Correspondances, où il est obligatoire de passer par ce recodage.
Le résultat est fondamental, il montre bien la "robustesse" de la règle de la majorité qui apparaît sous-jacente, même là où on ne l'attendait pas.

Propriété (4-2)
L'indice de DICE bis: Sdb que nous avons introduit ne serait lui aussi, si on l'applique au tableau mis sous forme disjonctive complète, qu'une variante de la règle de la majorité sur le tableau de données sous forme de partitions.
On a en effet:

$$Sdb(i,j) = \frac{2\,c_{ij}}{c_{ij} + m} = \frac{2Sd(i,j)}{Sd(i,j) + 1}$$

et:

$$c_{ij}(\text{codage partition}) = \frac{Sdb(i,j)}{2-Sdb(i,j)}$$

Ceci implique les résultats suivants:

$$c_{ij} \geq \frac{m}{2} \;==)\; Sdb \geq \frac{2}{3}$$

et:

$$Sdb \geq \frac{1}{2} \Longrightarrow c_{ij} \geq \frac{m}{3}$$

Propriétés (4-3)
L'indice de JACCARD Sj est une fonction homographique du terme général c_{ij} d'un tableau collectif C construit à partir d'un codage sous forme de partitions.
On obtient alors la fonction de c_{ij} suivante:

$$Sj(i,j) = \frac{c_{ij}}{2m - c_{ij}}$$

et:

$$c_{ij}\,(\text{codage partition}) = m\,\frac{2\,Sj(i,j)}{Sj(i,j) + 1}$$

Ceci implique le résultat suivant

$$cij \geq \frac{m}{2} \Longrightarrow Sj(i,j) \geq \frac{1}{3}$$

$$Sj(i,j) \geq \frac{1}{2} \Longrightarrow c_{ij} \geq \frac{2m}{3}$$

Propriété (4-4)
L'indice de SORENSEN Sso induit lui une règle de majorité au 4/5 ème

Remarque
En résumé,
- les indices de SOKAL-MICHENER, RAO1, DICE et MARCOTORCHINO-MICHAUD1 sont des indices directement liés à la règle majoritaire de CONDORCET, sous différents codages.

- d'autre part si l'on connait le nombre de variables m et les tableaux de similarités construits à partir des indices de SOKAL-MICHENER, SOKAL-SNEATH, ROGERS-TANIMOTO, RAO2, MARCOTORCHINO- MICHAUD2 etc... on peut directement obtenir, sans connaissance a priori du tableau de données brutes, les tableaux C collectifs correspondants et utiliser aussi les méthodes précédentes. Si l'on sait de plus que les tableaux de similarités ont été construits à partir de tableaux initiaux mis sous forme disjonctive complète, on peut rajouter à la liste précédente les indices de JACCARD-DICE et SORENSEN.

- enfin, comme nous l'avons vu précédemment, ces divers indices structurellement dissemblables et introduits historiquement à des moments différents ne sont, en fait, que des variations sur le thème de la règle de la majorité

de CONDORCET.

VIII MESURES DE LA QUALITE DES DONNEES ET DE LA VALIDITE DES RESULTATS OBTENUS.

Il faut distinguer ici deux notions tout à fait différentes:
-La notion de "Concordance" liée à la qualité des données
-La notion de "Cohérence" liée à la validité du résultat Y optimal obtenu.

8-1) MESURE DE LA QUALITE DES DONNEES

a) Indice de concordance généralisé

En généralisant le "tau" de KENDALL défini par ce dernier pour l'agrégation d'ordres totaux, on peut construire un indice mesurant l'accord entre les différents "juges" sur leurs opinions de préférences sur les n objets (tableaux Ck) ou entre différents "critères" par rapport aux similarités qu'ils ont induites sur les n objets, en utilisant les propriétés des tableaux C^k en posant:

$$I(\text{concordance}) = 1 - \frac{\sum_k \sum_{k'} d(C^k, C^{k'})}{m^2\, n(n-1)/2}$$

En effet cet indice vaudra 1 si tous les tableaux C^k sont identiques , car alors : $d(C^k, C^{k'}) = 0 \quad \forall\ k \text{ et } k'$

Compte tenu de la linéarisation des distances $d(C^k, C^{k'})$ en fonction des termes c_{ij} et c'_{ij} on peut montrer que :

$$I(\text{concordance}) = 1 - \frac{4 \sum_{i \neq j} c_{ij}(m - c_{ij})}{m^2\, n(n-1)}$$

Mais comme on l'a déjà vu, on a la relation suivante entre c_{ij} et c'_{ij}:

$$c_{ij} = (c'_{ij} + m)/2$$

D'où $\Sigma\, c_{ij}(m - c_{ij}) = 1/4\ \Sigma\, (m^2 - c'^2_{ij})$
et dès lors on a l'expression simplifiée de l'indice de concordance généralisé en fonction du tableau collectif C':

$$I(\text{concordance}) = \frac{\sum_{i \neq j} c'^2_{ij}}{m^2\, n(n-1)}$$

Remarque:
Dans le cas de l'agrégation des préférences si les données s'éloignent par trop d'une structure de "totalité", c'est à dire s'il y a eu beaucoup de non réponses, on peut définir l'indice modifié suivant:

$$I(\text{concordance}) = 1 - 4 \sum_{i \neq j} \frac{c_{ij}\, c_{ji}}{(c_{ij} + c_{ji})^2\, n(n-1)}$$

Cet indice est égal à l'indice généralisé si $c_{ij} + c_{ji} = m$

b) Propriété de l'indice de concordance généralisé

- L'indice généralisé vaut 0, lorsque $c_{ij} = 0 \ \forall\, i \neq j$ ce qui revient à $c_{ij} = m/2 = c_{ji}$, c'est à dire si le nombre de "critères" ou de "juges" pour lesquels iRj est égal au nombre de critères pour lesquels jRi. C'est le cas de l'ambiguité maximale pour la règle majoritaire de Condorcet.

- L'indice de concordance vaut 1 si:
$c_{ij} = m$ ou si $c_{ij} = -m \ \forall\, i \neq j$
c'est à dire si l'on se trouve dans le cas de "l'unanimité" (pour ou contre) par paires.

Remarque:
Cet indice a peu de chance de valoir 1 dans la pratique, en effet il renseigne uniquement sur "l'unanimité" des "juges" ou "critères" plutôt que sur leur "cohérence" interne. Nous allons éclairer cette dernière affirmation à l'aide d'un exemple.

Supposons que dans le cas d'un problème d'agrégation des préférences sur n objets (O1,O2,O3,...On) avec un nombre de juges impair soit m=2p+1, on ait la configuration suivante du tableau collectif C:

	O1	O2	O3	O4	....	On
O1		p+1	p+1	p+1	....	p+1
O2	p		p+1	p+1	....	p+1
O3	p	p		p+1	....	P+1
O4	p	p	p		p+1.	P+1
.	.	.	.	.	.	.
On	p	p	p	p	...	

Comme $c_{ij} = 1$ ou $-1 \ \forall\, i \neq j$
on a:

$$I(\text{concordance}) = 1/(2p+1)^2 = 1/m^2$$

donc si $m \rightarrow \infty$ $I(\text{concordance}) \rightarrow 0$

Or contrairement à ce que semble indiquer ce résultat, le

classement O1 , O2 , O3 .....On est très "cohérent", puisque pour tous les couples (i,j) relatifs à ce classement c'est à dire tels que i<j, une majorité stricte de juges a préféré i à j. N'ayant pas d'effet Condorcet collectif, on peut alors parler de "cohérence" maximale. Et dans ce cas précis, l'indice de "cohérence" généralisé que nous allons définir au prochain paragraphe vaudra 1.

8-2) MESURE DE LA QUALITE DU RESULTAT

a) Définition de l'indice de cohérence généralisé

Parallélement à la mesure de la qualité des données , on a la possibilité dans cette approche de définir un indice de "cohérence" généralisé. Cet indice ayant essentiellement pour but de mesurer l'adéquation du résultat obtenu Y^* (solution optimale des problèmes d'optimisation du chapitre V), a un résultat "potentiel idéal" que l'on chercherait à atteindre.

En effet la fonction que nous maximisons étant la fonction :

$$F(Y) = \sum_i \sum_j c'_{ij} \, Y_{ij}$$

les quantités Max F(Y) et Min F(Y), suivant les contraintes que nous imposerons à Y, peuvent prendre différentes valeurs. Pratiquement c'est la contrainte de transitivité qui empêche seule d'avoir une expression directe et explicite de ces valeurs du Maximum et Minimum de F(Y).

En effet, appelons

To:(l'ensemble des relations Y totales)
AS:(l'ensemble des relations Y antisymétriques)
To$\cap$AS:(l'ensemble des relations totales et antisymétriques)
S: (l'ensemble des relations Y symétriques)
Q: (l'ensemble des relations Y quelconques)

On a alors les valeurs suivantes de Max F(Y) pour Y

appartenant à l'un quelconque de ces ensembles (voir dans l'article de P.Michaud et F.Marcotorchino (21) le mode d'obtention de ces résultats).

1) $\underset{Y \in To}{\text{Max}} F(Y) = \sum_{i<j} \text{Max} \, (c'_{ij}, c'_{ji}, c'_{ij} + c'_{ji})$

2) $\underset{Y \in AS}{\text{Max}} F(Y) = \sum_{i<j} \text{Max} \, (0 , c'_{ij}, c'_{ji})$

3) $\underset{Y \in To \cap AS}{\text{Max}} F(Y) = \sum_{i<j} \text{Max} \, (c'_{ij}, c'_{ji})$

4) $\underset{Y \in S}{\text{Max}} F(Y) = \sum_{i<j} \text{Max} \, (0, c'_{ij} + c'_{ji})$

$$5)\ \underset{Y \in Q}{\text{Max}}\ F(Y) = \sum_{i<j} \text{Max}\ (0, c'_{ij}, c'_{ji}, c'_{ij} + c'_{ji})$$

Le principe de l'indice de "cohérence/incohérence" généralisé étant de mesurer la perte dans la valeur de Max F(Y) qu'impose à Y par rapport à une situation "potentielle idéale" d'être en plus "transitive". Ainsi si l'on désigne par "relation potentielle" le fait pour Y d'appartenir à l'un des ensembles To,AS,S etc.. définis précédemment, que l'on notera du terme générique "RP", l'indice d'incohérence généralisé mesurera l'écart entre le maximum potentiel pour Y ϵ RP et le maximum obtenu si l'on rajoute la contrainte de transitivité Y ϵ RP∩T, ainsi cet indice se définira formellement par:

$$I(\text{incohérence}) = \frac{\underset{Y\in RP}{\text{Max}}\ F(Y) - \underset{Y\in RP\cap T}{\text{Max}}\ F(Y)}{\underset{Y\in RP}{\text{Max}}\ F(Y) - \underset{Y\in RP}{\text{Min}}\ F(Y)}$$

et I(cohérence) = 1 - I(incohérence)

Ces deux indices varient de 0 à 1.

Donnons ici l'expression de cet indice d'incohérence généralisé dans le cas de problèmes d'agrégation des préférences et d'agrégation de similarités:

1) Cas de l'agrégation des préférences

Dans cecontexte la "relation potentielle" Yϵ RP sera une relation Y appartenant à l'ensemble AS∩To (ensemble des relations totales et antisymétriques) il suffit de remplacer RP par AS∩To et RP∩T par AS∩To∩T dans la formule précédente

On obtient alors l'expression suivante :

$$I(\text{incohérence}) = \frac{\sum_{i<j} \text{Max}\ (c'_{ij}, c'_{ji}) - \underset{Y\in AS\cap To\cap T}{\text{Max}}\ F(Y)}{\sum_{i<j} |c'_{ij} - c'_{ji}|}$$

2) Cas de l'agrégation de similarités

Dans ce cas il faut remplacer RP par S et RP∩T par S∩T dans la formule générale. On obtient alors la formule détaillée suivante:

$$I(\text{incohérence}) = \frac{\sum_{i<j} \text{Max}\ (0, c'_{ij} + c'_{ji}) - \underset{Y\in S\cap T}{\text{Max}}\ F(Y)}{\sum_{i<j} |c'_{ij} + c'_{ji}|}$$

b) Interprétation statistique des 2 indices précédents
(dans le cas de tableaux C' symétriques ou antisymétriques)

Par analogie avec la théorie des tests statistiques définissons par:

ERREUR de type I: le fait d'avoir dans la solution optimale Y^*

$Y_{ij} = 1$ alors que $c'_{ij} < 0$ (violation de la règle de la majorité)

ERREUR de type II: le fait d'avoir dans la solution optimale Y^*

$Y_{ij} = 0$ alors que $c'_{ij} \geq 0$ (violation de la règle de la majorité)

De la même façon on peut définir deux types d'adéquation. En effet on aura,

ADEQUATION de type I si $y_{ij} = 1$ et $c'_{ij} \geq 0$

ADEQUATION de type II si $y_{ij} = 0$ et $c'_{ij} < 0$

A partir de ces différents types de configurations on peut calculer les quantités suivantes:

N1 = Valeur absolue des erreurs de types I
P0 = Valeur des erreurs de type II
P1 = Valeur des adéquations de type I
N0 = Valeur absolue des adéquations de type II.

On a alors

$$N1 = -\sum_i \sum_j c'_{ij}\, y_{ij} \mid c'_{ij} < 0 \text{ et } y_{ij} = 1$$

$$P0 = \sum_i \sum_j c'_{ij}(1-y_{ij}) \mid c'_{ij} \geq 0 \text{ et } y_{ij} = 0$$

$$P1 = \sum_i \sum_j c'_{ij}\, y_{ij} \mid c'_{ij} \geq 0 \text{ et } y_{ij} = 1$$

$$N0 = -\sum_i \sum_j c'_{ij}\,(1 - y_{ij}) \mid c'_{ij} < 0 \text{ et } y_{ij} = 0$$

Formons alors le tableau de contingence 2X2 suivant croisant les valeurs y_{ij} de Y et les valeurs c'_{ij} de C':

c'_{ij} \ y_{ij}	1	0
≥ 0	P1	P0
< 0	N1	N0

Il semble intuitif à partir de ce tableau de considérer que la somme N1 + P0 représente bien la notion d'incohérence; plus cette somme est faible plus la cohérence est parfaite.

En effet on peut montrer facilement que si les tableaux C' sont antisymétriques (agrégation des préférences) ou symétriques (agrégation de similarités) les indices d'incohérence définis en 1) et 2) du paragraphe précédent s'expriment en fonction des quantités N1,N0,P1,P0 selon la formule très simple suivante:

$$I(\text{incohérence}) = \frac{N1 + P0}{N1 + P0 + P1 + N0}$$

Remarque:

A partir du tableau de contingence croisant Y et C' on peut calculer soit un χ^2 soit le coefficient de correlation mutuelle associé (qui lui est équivalent à une constante près).

On obtient:

$$\rho^2 = \frac{(P1.N0 - N1.P0)^2}{(P1+P0)(N1+N0)(P1+N1)(P0+N0)}$$

Si le tableau C' est antisymétrique on a P1 = N0 et N1 = P0 d'où l'indice d'incohérence se ramène a :

I(incohérence) = N1/(N1+P1)et I(cohérence) = P1/(P1+N1)

Mais justement dans ce cas $\rho^2 = (P1-N1)^2/(P1+N1)^2$

et on a alors (dans le cas de l'agrégation des préférences)

$$\rho = I(\text{cohérence}) - I(\text{incohérence})$$

Dans le cas où les tableaux C' ne sont pas antisymétriques si l'on se trouve en "agrégation des préférences" ou symétriques si l'on se trouve en "agrégation des similarités", on peut transformer les tableaux C' de départ en tableaux C" régularisés. C'est à dire en posant $c''_{ij} = c'_{ij} - c'_{ji}$ dans le premier cas ou en posant $c''_{ij} = c'_{ij} + c'_{ji}$ dans le second cas.

8-3) MESURE DE L'HOMOGENEITE DES CLASSES (en agrégation des similarités)

Là encore on peut mesurer d'une part l'homogénéité en concordance d'autre part l'homogénéité en cohérence, définissant de ce fait deux indices différents.

a) Homgénéité en concordance

Cet indice mesure le fait que les objets contenus dans une classe ont été considérés comme se ressemblant deux à deux pour une grande quantité de critères. On le définit par:

$$I(\text{homogénéité conc}) = \frac{1}{p} \sum_{k=1}^{p} \sum_{\substack{i,j \in H_k \\ i \neq j}} \frac{c'_{ij}}{m\, n_k(n_k-1)}$$

Avec p = nombre de classes de la partition optimale

n_k = cardinal de la classe numéro k : (H_k)

Cet indice varie entre 0 et 1, il vaudra 1 si $c'_{ij} = m$ pour tous les couples (i,j) appartenant à une même classe.

b) Homogénéité en cohérence

Cet indice s'obtient en remplaçant c'_{ij} par Sign (c'_{ij}) et m par 1 dans la formule précédente. Cet indice vaudra 1 s'il n'y a pas d'erreur de typeI dans les différentes classes formées.

IX - EXEMPLES D'APPLICATION

Afin d'illustrer l'ensemble de la méthodologie proposée dans les chapitres précédents, nous avons choisi de présenter des exemples "d'agrégation des similarités" plutôt que "d'agrégation des préférences", dans la mesure où le lecteur pourra trouver dans (16) et dans (21) bon nombre d'exemples "d'agrégation des préférences" amplement détaillés.

Nous présenterons ici trois exemples:

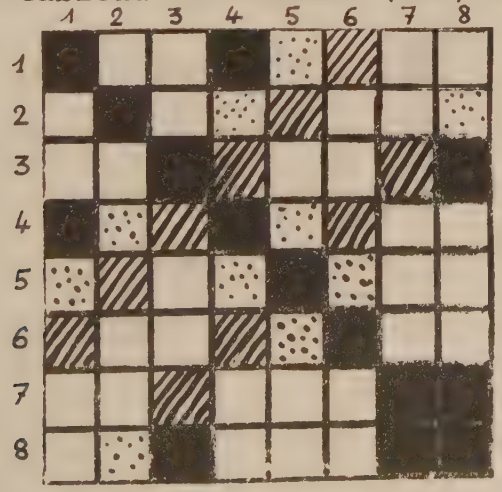
- le premier exemple (purement illustratif) aura pour thème les tableaux de noircissement de BERTIN-CZEKANOWSKI;

- le deuxième exemple (réel) traitera de la classification des félidés suivant des caractéristiques morphologiques externes ou internes, et des caractéristiques de comportement telles que J. DORST les présente dans (10);

- le troisième exemple (également réel) aura pour thème la classification de certains pays de l'ONU suivant leurs attitudes de vote aux motions proposées en 1968 (ces données ont été extraites du livre (12) de M. JAMBU).

9-1) TRAITEMENT DES TABLEAUX DU TYPE BERTIN-CZEKANOWSKI

Considérons le tableau de données (8X8) suivant:



La case (i,j) peut prendre les quatre valeurs de "grisé" suivantes:

En fonction du niveau de "grisé" de la case (i,j) la "liaison" ou la "ressemblance" entre i et j sera plus ou moins forte:

Ainsi ■ signifie que i et j sont très fortement "liés" ou "ressemblants";

▨ signifie que i et j sont "liés" ou "ressemblants";

▩ signifie que i et j sont peu "liés" ou peu "ressemblants";

□ signifie que i et j ne sont pas "liés" ou pas "ressemblants".

Ces tableaux où les cases (i,j) prennent des valeurs de "grisé" plus ou moins fortes sont sourtout utilisés en psychologie ou en botanique. L'idée sous-jacente étant de ramener le long de la diagonale principale de ce tableau carré symétrique les valeurs les plus noires, afin de définir une partition ou des zônes de ressemblances de la population étudiée.
Ce genre de tableaux relève de traitements par des techniques astucieuses pseudo-manuelles de permutations simultanées de lignes et colonnes (BERTIN); ou, après recodage, de traitements par analyse factorielle des correspondances, en espérant soit la présence d'un effet en "U" (effet Guttman) donnant une indication, par projection des points sur l'axe 1, sur une "bonne" permutation des lignes et des colonnes, soit des regroupements par paquets suivant les axes 1 et 2.

Nous montrons ici que la méthode présentée en IV-2) s'adapte parfaitement au traitement de tels tableaux.
En effet, considéreons pour notre exemple la correspondance suivante entre le "grisé" et le tableau C collectif (voir 7-1-b-iii):

c_{ij} = 3 si valeur case (i,j) = ■

c_{ij} = 2 si valeur case (i,j) = ▨

c_{ij} = 1 si valeur case (i,j) = ▩

c_{ij} = 0 si valeur case (i,j) = □

On a donné la valeur 3 à cii par homogénéité; mais ces

valeurs ne joueront aucun rôle dans les calculs.

Alors en posant C' = 2C-3 (voir justification en 7-1-b-iii), on obtient le tableau C' collectif suivant:

	1	2	3	4	5	6	7	8
1	3	-3	-3	3	-1	1	-3	-3
2	-3	3	-3	-1	1	-3	-3	-1
3	-3	-3	3	1	-3	-3	1	3
4	3	-1	1	3	-1	1	-3	-3
5	-1	1	-3	-1	3	-1	-3	-3
6	1	-3	-3	1	-1	3	-3	-3
7	-3	-3	1	-3	-3	-3	3	3
8	-3	-1	3	-3	-3	-3	3	3

La maximisation de $F(Y) = \sum_i \sum_j c'_{ij} y_{ij}$

sous les contraintes:

$$y_{ij} - y_{ji} = 0$$

$$y_{ij} + y_{jk} - y_{ik} \leq 1$$

$$y_{ij} \in \{0,1\}$$

donne ici le résultat suivant:

Max F(Y) = 26 (la diagonale n'étant pas considérée) avec:

Y^*= Classe 1 = 1,4,6

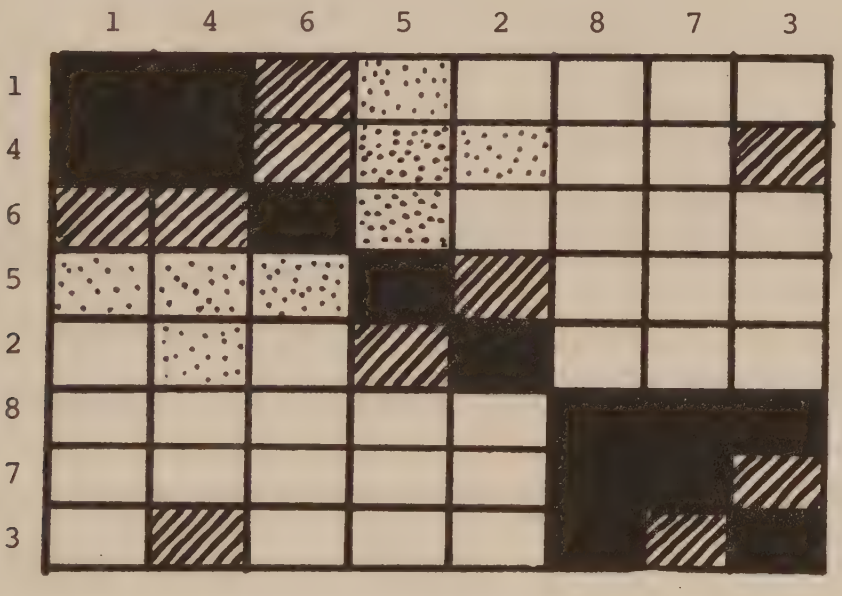
Classe 2 = 5,2

Classe 3 = 8,7,3

On obtient alors le nouveau tableau C' permuté sur la relation d'équivalence optimale Y^* précédente, soit:

	1	4	6	5	2	8	7	3
1	3	3	1	-1	-3	-3	-3	-3
4	3	3	1	-1	-1	-3	-3	1
6	1	1	3	-1	-3	-3	-3	-3
5	-1	-1	-1	3	1	-3	-3	-3
2	-3	-1	-3	1	3	-3	-3	-3
8	-3	-3	-3	-3	-3	3	3	3
7	-3	-3	-3	-3	-3	3	3	1
3	-3	1	-3	-3	-3	3	1	3

Ce qui en redonnant aux valeurs C' leurs correspondances en "grisé" nous permet d'obtenir la configuration du tableau de données suivante:



Il est intéressant de mesurer ici les valeurs des indicateurs de "concordance" et de "cohérence":

$$I(\text{concordance}) = \frac{\sum_{i \neq j} c'^{2}_{ij}}{m^2\, n(n-1)} = \frac{360}{504} = 0{,}714$$

Cette "concordance" très forte (compte tenu de ce qui a été dit en 8-1), nous indique que le traitement que l'on a fait subir aux données était judicieux.

De même on a:

$$I(\text{incohérence}) = \frac{\sum_{i<j} \text{Max}(0, c'_{ij} + c'_{ji}) - 26}{\sum_{i<j} |c'_{ij} + c'_{ji}|} = \frac{2}{80}$$

et:

$$I(\text{cohérence}) = \frac{78}{80} = 0,975$$

Le ρ^2 calculé sur le tableau de contingence:

C'	Y = 1	Y = 0
≥ 0	26	1
< 0	0	104

$$\rho^2 = \frac{(26 \times 104)^2}{28 \times 104 \times 26 \times 106} = \frac{26 \times 104}{28 \times 106} = 0,9110$$

et:

$$\rho = 0,954$$

Donc le résultat obtenu montre que les classes sont très cohérentes. C'est à dire qu'il y a peu de classes empiétantes ou non homogènes.

L'homogénéité en concordance = 0,55

L'homogénéité en cohérence = 1

9-2) APPLICATION A LA CLASSIFICATION DES FELIDES

a) Présentation du problème

Il s'agit dans cet exemple de classifier 30 félins provenant des trois continents: AFRIQUE, EURASIE, AMERIQUE; à l'aide de 14 paramètres (ou caractéristiques) morphologiques ou de comportement.

Les 30 félins se répartissent comme suit:

AFRIQUE	EURASIE	AMERIQUE
Lion	Tigre	Jaguar
Léopard	Once	Puma
Guépard	Léopard	Ocelot

Serval	Pant. Nébul.	Jaguarundi
Caracal	Lynx	Ch. Marguay
Ch.Viverrin	Ch. de Chine	Ch. Tigrin
Ch. Chaus	Ch. du Bengale	Ch. des Andes
Ch. Doré	Ch. Rouilleux	Ch. Marbré
Ch. Marguerite	Ch. Malais	
Ch. Cafer	Ch. de Bornéo	
Ch. Nigripes	Ch. Manul	
	Ch. Temminck	

Ces différents félins sont de genre et d'allure différents (voir certains d'entre-eux sur la planche N° 1).
Certains sont du genre panthera, d'autres neofelis, d'autres felis, un du genre acynonix; leurs poids varient de 5 à 150 kilos, etc...
Les différents paramètres choisis ont été extraits de la liste descriptive généralement utilisée par J. DORST pour décrire une espèce animale ou une variété dans son livre (10).
Ces différents paramètres se répartissent comme suit:

CODE	PARAMETRES	MODALITES
	1) Paramètres morphologiques	
1	(TYPPEL) Aspect du pelage	1 Sans taches, uni 2 Tacheté 3 Rayé 4 Marbré
2	(LONG POIL) Fourrure	0 Poils ras 1 Poils longs
3	(OREILLES) Oreilles	1 Ronde ou arrondie 2 En pointe (type caracal)
4	(TAILLE) Taille au garrot	1 Taille ≤ 50 cm. 2 50 cm. <taille ≤70 cm. 3 Taille > 70 cm.
5	(POIDS) Poids	1 Poids ≤ 10kg. 2 10 kg.< poids ≤ 80kg. 3 Poids > 80kg.
6	(LONGUEUR) Longueur du corps	1 Longueur ≤ 80cm. 2 80cm. <longueur ≤150cm. 3 Longueur> 150cm.
7	(QUEUE) Longueur de la queue relativement à la longueur du corps	1 Petite 2 Moyenne 3 Longue

PUMA

felis concolor

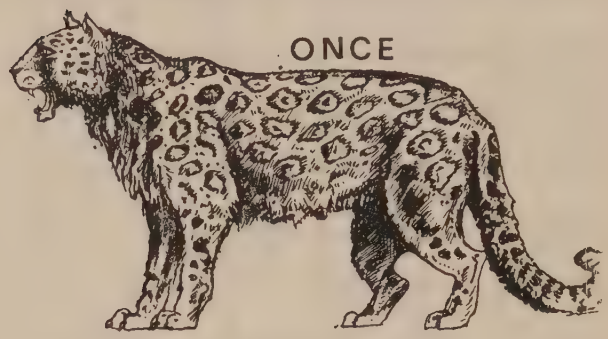
JAGUAR

panthera onca

GUEPARD

acinonyx jubatus

ONCE

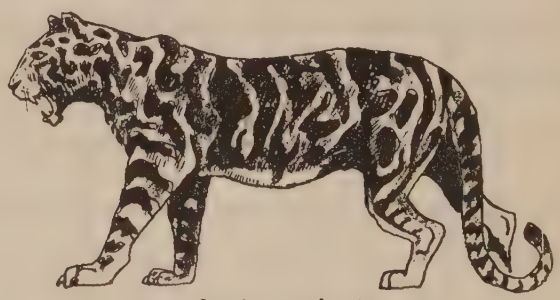


panthera uncia

CARACAL

felis caracal

PANTHERE NEBULEUSE



neofelis nebulosa

CHAT DE CHINE

felis bieti

SERVAL

felis serval

PLANCHE N°1

jfm (1980)

	TYPPEL	LONGPOIL	RETRACT	COMPORT	OREILLES	LARYNX	TAILLE	POIDS	LONGUEUR	QUEUE	DENTS	TYPPROIE	ARBRES	CHASSE
LION	1	0	1	1	1	1	3	3	3	2	1	1	0	1
TIGRE	3	0	1	3	1	1	3	3	3	2	1	1	0	0
JAGUAR	2	0	1	2	1	1	3	3	2	1	1	1	1	0
LEOPARD	2	0	1	3	1	1	3	3	2	2	1	2	1	0
ONCE	2	1	1	1	1	1	2	2	2	3	1	2	1	0
GUEPARD	2	0	0	1	1	0	3	2	2	3	0	2	0	1
PUMA	1	0	1	2	1	0	2	3	2	3	1	2	1	0
NEBUL	4	0	1	3	1	1	2	2	2	3	1	3	1	0
SERVAL	2	0	1	1	2	0	2	2	2	1	0	3	1	1
OCELOT	2	0	1	2	1	0	2	2	2	2	0	3	1	0
LYNX	2	1	1	2	2	0	2	2	2	1	1	2	1	0
CARACAL	1	0	1	2	2	0	2	2	1	1	0	3	1	1
VIVERRIN	2	0	1	2	1	0	1	1	2	2	0	3	0	0
YAGUARUN	1	0	1	2	1	0	1	2	2	3	0	3	1	0
CHAUS	1	1	1	3	2	0	1	2	1	2	0	3	1	0
DORE	1	0	1	3	1	0	1	1	1	2	0	3	1	0
MERGUAY	2	0	1	3	1	0	1	1	1	2	0	3	1	0
MARGERIT	1	1	1	2	1	0	1	1	1	2	0	3	0	0
CAFER	3	0	1	3	1	0	1	1	1	2	0	3	1	1
CHINE	1	0	1	2	2	0	1	1	1	1	0	3	1	0
BENGALE	2	0	1	3	1	0	1	1	1	2	0	3	1	0
ROUILLEU	2	0	1	2	1	0	1	1	1	2	0	3	1	0
MALAIS	1	1	1	3	1	0	1	1	1	1	0	3	1	0
BORNEO	1	0	1	3	1	0	1	1	1	2	0	3	1	0
NIGRIPES	2	0	1	2	1	0	1	1	1	1	0	3	1	1
MANUL	1	1	1	3	1	0	1	1	1	1	0	3	1	0
MARBRE	4	0	1	3	1	0	1	1	1	3	0	3	1	0
TIGRIN	2	0	1	3	1	0	1	1	1	2	0	3	1	0
TEMMINCK	1	0	1	3	1	0	1	1	1	2	0	3	1	0
ANDES	2	1	1	3	1	0	1	1	2	2	0	2	1	0

8 (DENTS)	Canines développées	1 Très développées 2 Peu développées
9 (LARYNX)	Présence ou non de l'os hyaoide	1 Présence 2 Absence
10 (RETRACT)	Griffes rétractiles	1 Oui 2 Non

2) Paramètres de comportement

11 (COMPORT)	Comportement prédateur	1 Diurne 2 Diurne et nocturne 3 Nocturne
12 (TYPPROIE)	Type de la proie	1 Grosse proie (antilope, buffle, etc.) 2 Grosse ou petite proie 3 Petite proie (musaraigne, petit singe)
13 (ARBRES)	Monte ou non aux arbres	1 Oui 2 Non
14 (CHASSE)	Chasse à courre ou à l'affut	1 Oui 2 Non

A partir de ces variables on obtient le tableau de données (30X14) reproduit en page 57.

Chacune des 14 variables déterminant une partition de l'ensemble des félins en autant de classes que la variable a de modalités. La recherche de la "partition centrale" consistera à chercher le Maximum de F(Y) sous des contraintes de transitivité et de symétrie définies au Chapitre IV.

b) Résultats

On a alors le résultat suivant:

$$\text{Max } F(Y) = F(Y^*) = 2608$$

Y* correspondant à la partition solution:

Classe 1 : LION, TIGRE.

Classe 2 : JAGUAR, LEOPARD, ONCE, PUMA, PANT.NEBUL,

LYNX.

Classe 3 : SERVAL, CARACAL, OCELOT, VIVERRIN, JAGUARUNDI, CHAUS, DORE, MARGUAY, MARGERIT, CAFER, CHINE, BENGALE, ROUILLEUX, MALAIS, BORNEO, NIGRIPES, MANUL, MARBRE, TIGRIN, TEMMINCK, ANDES.

Classe 4 : GUEPARD.

I(cohérence) = 0,9543 , I(homogénéite) = 0,8443

Cette partition optimale donne un résultat en accord avec les connaissances généralement admises par les zoologues sur la classification des félins par genres et par espèces.

En effet:
La Classe 1 regroupe les 2 plus gros félins (le tigre et le lion) (genre panthera), que leur corpulence et leur comportement distinguent de façon évidente de l'ensemble des autres.

La Classe 2 (classe d'intermédiarité), regroupe d'une part les 3 plus petits félins du genre "panthera" (Jaguar, leopard,once) d'autre part les 2 plus grand félins du genre "felis" (le puma et le lynx) ainsi qu'un félin d'un genre intermédiaire (la panthère nébuleuse du genre "neofelis"). Cette classe caractérise donc les félins de taille intermédiaire entre les "grands" et les "petits" félins. Ceci d'autant plus qu'il existe des variations intraspécifiques pour certains des félins de cette classe compliquant la difficulté d'affectation à une classe précise,(ainsi le leopard des savanes africaines peut-il peser plus de 60 kg et être considéré alors comme un félin de grande taille alors que son homologue des forêts asiatiques ne pèse lui qu'une vingtaine de kilos).

La Classe 3 caractérise les petits félins, c'est à dire à l'exception du puma et du lynx tous les félins du genre "felis". Ce sont en général des espèces que l'on peut schématiquement affubler du nom de "chat x...".

La Classe 4 enfin contient le guépard tout seul. Ce résultat est très satisfaisant du point de vue de la classification zoologique des félins, car pour la plupart des zoologues le guépard ne devrait pas être considéré comme un félin. Il est en effet morphologiquement très différents des autres félidés; d'ailleurs son nom latin d'acinonyx jubatus traduit cette ambiguité.

9-3) APPLICATION A L'ANALYSE DE SCRUTINS A L'ONU EN 1968

a) Présentation du problème

Pour cet exemple les données ont été extraites du livre de M.Jambu "Classification automatique pour l'analyse des données, Tome II pp 38-41 (12). Bien qu'incomplètes comme le signale M.Jambu p.38, elle nous servirons néanmoins à illustrer la méthode sur des données totalement qualitatives, sans aucune notion de hierarchie sous-jacente (contrairement au cas de certains des parametres de l'exemple précédent comme le poids ou la taille).

Il s'agit des votes de 54 pays à l'assemblée des Nations Unies en 1968 sur 3 motions, correspondant au tableau de données brutes suivant:

U.S.	3	3	1
CANA	3	2	1
CUBA	1	5	3
HAIT	5	1	1
DOMI	1	1	1
JAMA	3	1	1
TRIN	3	1	1
BARB	5	1	2
MEXI	3	1	1
GUAT	3	1	1
HOND	3	2	1
EL S	3	2	1
NICA	3	2	1
COST	3	1	1
PANA	3	1	1
COLU	5	2	1
VENE	3	1	1
GUYA	5	1	1
ECUA	3	5	2
PERU	3	1	1
BRAZ	3	2	1
BOLI	5	1	1
PARA	3	2	1
CHIL	3	1	1
ARGE	3	1	1
URUG	3	5	1
U.K.	3	3	1
IREL	3	2	1
NETH	3	3	1
BELG	3	3	1
LUXE	3	3	1
FRAN	3	2	3
SPAI	3	2	1
PORT	5	3	2
POLA	1	1	3
AUST	3	2	2
HUNG	1	1	3
CZEC	1	1	3
ITAL	3	2	1
MALT	5	5	1
ALBA	1	5	3
YUGO	1	1	3
GREE	3	1	1
CYPR	3	1	1
BULG	1	1	3
ROMA	1	1	3
USSR	1	1	3
UKRA	1	1	3
BYEL	1	1	3
FINL	2	2	3
SWED	3	2	3
NORW	3	2	3
DENM	3	2	3
ICEL	3	2	1

Il y a 4 modalités par scrutin: 1=OUI , 2=NON, 3=ABSTENTION, 5=ABSENCE

Chacun des votes détermine donc une partition des 54 pays en au plus 4 classes, correspondant aux 4 modalités possibles pour chaque scrutin.

On recherche ici une partition des 54 pays maximisant la fonction F(Y) sur les contraintes definies en 4-2-b).

b) Résultats

La valeur Max F(Y)= F(Y*)= 1596

Y* induit la partition optimale en 6 classes, suivante:

Classe 1: USA, CANADA, JAMAIQUE, TRINIDAD, MEXIQUE, GUATEMALA, HONDURAS, SALVADOR, NICARAGUA, PANAMA, VENEZUELA, PEROU,BRESIL, PARAGUAY, CHILI, ARGENTINE, URUGUAY, U.K., IRLANDE, PAYS-BAS, BELGIQUE, LUXEMBOURG, ESPAGNE, ITALIE, GRECE, CHYPRE, ISLANDE.

Classe 2:CUBA, DOMINIQUE, POLOGNE, HONGRIE, TCHECOSLOVAQUIE, ALBANIE, YOUGOSLAVIE, BULGARIE, ROUMANIE, URSS, UKRAINE, BIELORUSSIE.

Classe 3: HAITI, BARBADES, COLOMBIE, GUYANA, BOLIVIE, MALTE.

Classe 4: FRANCE, AUTRICHE, FINLANDE, SUEDE, NORVEGE, DANEMARK.

Classe 5: EQUATEUR.

Classe 6: PORTUGAL.

I(coherence)= o,9517 I(concordance)= 0,440

Compte tenu de la liste des pays considérée ici, cette partition se révèle très proche de l'idée du profil politique que l'on pouvait se faire de ces 54 pays en 1968.

En effet, la classe 1 regroupe l'ensemble des pays alignés sur la politique des Etats-Unis: 14 pays d'Amérique Latine, 10 pays européens et 3 pays nord américains.

La classe 2 regroupe l'ensemble des pays politiquement alignés sur l'URSS: 10 pays d'Europe de l'Est et Cuba plus la Dominique.

La classe 4 regroupe ce qu'en 1968 (présence de De Gaulle) on pouvait appeler les pays neutres ou non alignés Européens; c'est à dire les Pays Scandinaves, la France et l'Autriche.

La classe 3 ne s'interprète pas très clairement, mais neanmoins on constate la présence de 3 pays insulaires (Malte, les Barbades et Haiti) et de 3 pays politiquement à "droite" d'Amérique Latine.

Les classes 5 et 6 contenant le Portugal et l'Equateur

seulement ne sont pas directement interprétables.

Remarque:
Il faut noter qu'ici ce résultat a été obtenu d'une part sans fixation a priori du nombre de classes, d'autre part sans transformation du tableau des données brutes, en particulier sans être obligé de passer par une mise sous forme disjonctive complète.

XI- REFERENCES

(1) K.ARROW, Social Choice and individual value, Wiley, New-York, (1963).

(2) M.BARBUT,Note sur les ordres totaux à distance minimum d'une relation binaire donnée, Math. et Sciences humaines N°17, pp. 47-48, (1966).

(3) J.P.BARTHELEMY, Sur les éloignements symétrique et le principe de Pareto, Math. et Sciences Humaines, N°56,pp. 97-125, (1976).

(4) J.P.BARTHELEMY et B.MONJARDET, Ajustement et Résumé de données relationnelles: les relations centrales, in Data Analysis and Informatics, North-Holland, (1980).

(5) J.P.BENZECRI et coll, L'Analyse des Données, Tome I: la Taxinomie, Tome II: l'Analyse des Correspondances, Paris, Dunod, (1973).

(6) D.BLACK,The Theory of Comittees and Elections, Cambridge University Press, Cambridge (1958).

(7) J. de CANI, Maximum likelihood paired comparison ranking by linear programming, Biometrika, Vol.56, N°3, 537, (1969).

(8) J. de CANI, A branch and bound algorithm for maximum likelihood paired comparison ranking, Biometrika, Vol.59, N°1, pp.131-135, (1972).

(9) E.DIDAY, Nouvelle méthode et nouveaux concepts en Classification Automatique et Reconnaissance des Formes, Thèse de Doctorat d'Etat, Université Paris VI, (1972).

(10) J.DORST et P.DANDELOT, Guide des grands mammifères

d'Afrique, Delachaux et Niestlé, Neuchâtel, (1972).

(11) D.M.JACKSON, Classification, Relevance, and Information Retrieval,in Advances in Computers Edité par Franz L.Alt and al. Vol 11, Academic Press, New York, (1971).

(12) M.JAMBU, Classification automatique pour l'analyse des données. Tome I: méthodes et algorithmes, avec M.O.LEBEAUX Tome II: Logiciels, Dunod (1978).

(13) J.KEMENY et L.SNELL, Mathematical Models in the Social Sciences, The M.I.T. Press, Cambridge U.S.A., (1962).

(14) M.G.KENDALL, Rank Correlation Methods, Griffin, Londres, (1962).

(15) I.C.LERMAN, Les Bases de La Classification Automatique, Gauthier Villars, Paris, (1970).

(16) F.MARCOTORCHINO et P.MICHAUD, Optimisation en Analyse Ordinale des Données, Masson, Paris, (1979).

(17) F.MARCOTORCHINO et P.MICHAUD, Optimization in exploratory Data Analysis, in the "proceedings" du Vème Symposium International de Rech. Oper., Köln, (Publié par Physika Verlag 1981). Disponible en version augmentée comme Etude F03 Centre Scientifique IBM-France Paris,(1980).

(18) P.MICHAUD et F.MARCOTORCHINO, Optimization in Ordinal Data Analysis, IBM-France Scientific Center Technical Report 001, (1978). (Une partie importante de ce papier a été présentée au Symposium International d'Optimisation Discrète, Vancouver, Août 1977).

(19) P.MICHAUD et F.MARCOTORCHINO, Optimisation en Analyse des Données Relationnelles, Data Analysis and Informatics, E.Diday and all. editeurs, North Holland, (1980).

(20) P.MICHAUD et F.MARCOTORCHINO, Modèles d'Optimisation en analyse des données relationnelles, Math. et Sciences Humaines N°67, pp. 7-38, Dunod, Paris, (1979).

(21) P.MICHAUD et F.MARCOTORCHINO, Une "nouvelle" méthode d'agrégation d'opinions: application à une enquête de l'INSEE,(à paraitre dans les Annales de l'INSEE).

(22) B.MIRKIN, Group Choice, édité par P. Fichburn, V.Winston and Sons, J.Wiley, (1979).

(23) B.MONJARDET, Tournois et ordres médians pour une opinion, Math. et Sciences Humaines N°43, pp. 55-70, (1973).

(24) S.REGNIER, Sur Quelques Aspects Mathématiques des Problèmes de Classification Automatique, I.C.C. Bulletin, Rome, (1965).

(25) C.T.ZAHN, Approximating Symmetric Relations by Equivalence Relations, SIAM J. of Applied Math. N°12, p 840, New York, (1964).

T2

OPTIMISATION EN ANALYSE DES DONNEES RELATIONNELLES

Pierre Michaud
François Marcotorchino

Développement scientifique IBM France

1. INTRODUCTION

Aujourd'hui l'ensemble des méthodes d'analyse des données est extrêmement important; néanmoins il semble exister des domaines où la connaissance théorique des problèmes est en avance sur les possibilités de résolution pratique. L'agrégation des préférences et la recherche de partitions optimales en classification automatique en sont deux bons exemples.

La connaissance théorique et axiomatique de ces problèmes ne date pas d'aujourd'hui; le problème du paradoxe des votes, abordé par Antoine de Condorcet, date de 1785. En revanche les algorithmes de résolution sont extrêmement limités.

Dans le cas de l'agrégation des préférences il n'existe à l'heure actuelle que des méthodes d'énumération implicite pour résoudre le problème de façon exacte, or ces méthodes sont très onéreuses en temps de calcul et sont limitées à des tailles n'exédant pas n=20. Dans le cas de la "partition optimale", il existe des algorithmes performants, comme la méthode des "nuées dynamiques" de E.Diday, mais cette méthode, comme les autres n'est qu'une méthode heuristique ne garantissant qu'un optimum local du problème considéré.

le propos de la présente communication est de montrer que pour ces problèmes, et d'autres problèmes voisins, il existe une modélisation générale, conduisant à l'élaboration d'algorithmes performants donnant une solution exacte aux problèmes evoqués ci-dessus.

Le point historique et structurel étant fait dans la communication (2) "Ajustement et Résumés de données Relationnelles : Les Relations Centrales",présentée par M.M Barthélémy et Monjardet à ce même symposium, nous insisterons ici sur les aspects de modélisation, d'algorithmique, et d'applications.

Les techniques utilisées reposeront essentiellement sur la programmation linéaire, les algorithmes associés faisant appel à des systèmes informatiques développés en PL/1, ainsi qu'au système de programmation mathématique MPSX/370 possédant un langage de contrôle (ECL) en PL/1.

2. POSITION DU PROBLEME GENERAL

2.1. PRESENTATION DU PROBLEME

Le problème général que nous traiterons dans la présente communication est le suivant: il s'agit de "l'agrégation" ou de "l'approximation" de plusieurs relations binaires par une relation binaire particulière.
les problèmes évoqués dans l'introduction, tels que le problème de l'agrégation des préférences ou le problème de la recherche d'une "Partition Centrale", en sont des cas particuliers.

Dans le premier de ces deux cas, un certain nombre de relations binaires sont à approximer par une relation d'ordre total. Dans le second cas elles sont à approximer par une relation d'équivalence.
Les relations d'ordre total et d'équivalence ont en commun la propriété suivante: ce sont des relations transitives particulières.
Mais il existe d'autres types de relations transitives.

Les relations transitives peuvent être divisées en six groupes:

1) Une relation transitive peut n'être que transitive, c'est alors une relation de préordre.

2) Lorsqu'une relation transitive (notée $\leq$) est en même temps anti-symétrique, c'est à dire que $i \leq j$ implique $j \$ i$,

c'est une relation d'ordre.

3) Lorsqu'une relation transitive est en même temps symétrique, c'est à dire que $i \leq j$ implique $j \leq i$, c'est une relation d'équivalence.

Pour chacun de ces trois cas, il y a encore deux possibilités; la relation peut être soit "partielle" soit "totale".
Une relation est dite "partielle " dans le cas général .
Une relation est dite "totale", lorsqu'au moins l'une des deux éventualités $i \leq j$ ou $j \leq i$, se produit pour tout couple (i,j) avec $i \neq j$.

L'un de ces groupes, le groupe des relations d'équivalence totales, ne présente aucun intérêt pratique, et ne sera pas considéré ici; il correspond à l'unique configuration où tous les éléments sont équivalents et appartiennent à une même classe.

Parmi les cinq cas restants, deux sont particulièrement intéressants au point de vue pratique , il s'agit des relations d' ordre total et d'équivalence, que nous étudierons en détail par la suite; les autres cas sont les suivants:
-Relation de préordre total.
-Relation de préordre partiel.
-Relation d'ordre partiel.

Le problème général de l'agrégation de relations binaires, que nous considérons, consiste précisemment à approximer des relations binaires par l'une ou l'autre de ces relations transitives.
L'approximation de m relations binaires données par l'un de ces cinq types de relations transitives sera réalisée par une approche générale valable, à quelques modifications près, pour l'ensemble des cas que nous avons considérés précédemment.

2.2. PRESENTATION DES DONNEES

Chacune des m relations binaires de départ, définissant les données individuelles d'un "juge" ou "critère" numéro k, sera représentée par un tableau C^k. Ce tableau, que nous appellerons tableau d'opinions ou de comparaisons, est un tableau en (0,1) de terme général:

$c^k_{ij} = 1$ Si i "est en relation" avec j
c'est à dire si iRj.
$c^k_{ij} = 0$ Sinon

Par convention la valeur c^k_{ij} avec i=j ne sera jamais

considérée.

Suivant le type de problème à traiter et la nature des données à considérer, la signification des valeurs c sera différente.

Par exemple:
-i est "en relation avec " j signifie : i est "préféré à " j s'il s'agit d'un tableau de préférences.

-i est "en relation avec " j signifie : i est "semblable à " j s'il s'agit d'un tableau de similarités.

A chaque relation binaire représentée par un tableau C^k, on associe un autre tableau, noté C'^k ; ce nouveau tableau, à valeurs (-1,+1), est défini de la façon suivante:

$c'^k_{ij} = 1$ si l'objet i "est en relation " avec l'objet j, pour le "juge" k; c-à-d si iRj.
$c'^k_{ij} = -1$ sinon

l'expression de c'^k_{ij} en termes de c^k_{ij} est donnée par

$$c'^k_{ij} = 2c^k_{ij} - 1$$

L'obtention des tableaux C^k ou C'^k peut se faire directement ou par l'intermédiaire de tableaux de données annexes.
Exemples:
- Dans le cas où les données brutes se présentent sous forme de tableaux de rangs r_{ik} (rang de l'objet i pour le juge k) ou de notes n_{ik} (note du juge k pour l'objet i), le tableau C^k associé s'obtient de la façon suivante:

$c^k_{ij} = 1$ si $n_{ik} \geq n_{jk}$ ou si $r_{ik} \leq r_{jk}$.
$c^k_{ij} = 0$ sinon

- Dans le cas où les données représentent des tableaux de proximités Pk, (calculées à partir de distances, d'indices de dissimilarités ou de mesures de proximité non nécessairement symétriques), le tableau C associé sera obtenu de la façon suivante:

$c^k_{ij} = 1$ si p^k_{ij} est inférieur à un seuil donné
$c^k_{ij} = 0$ sinon

En sommant tous les tableaux individuels C^k, représentant les relations binaires de départ, on obtient un tableau collectif C, qui représente l'opinion de l'ensemble des m "juges".

On a donc : $C = \sum_k C^k$ et $c_{ij} = \sum_k c^k_{ij}$

Le terme général cij du tableau collectif C représente d'une façon générale le nombre de relations binaires où i est en relation avec j; c'est le nombre de juges ayant mis en relation i avec j.

De la même façon le tableau collectif C'est défini par:

$$C' = \sum_k C'^k \quad \text{et} \quad c'_{ij} = \sum_k c'^k_{ij} .$$

Le terme général c'ij du tableau collectif C' est égal au nombre de juges ayant mis i en relation avec j moins le nombre de juges n'ayant pas mis i en relation avec j.

2.3. DEFINITION DU PROBLEME ET PROPRIETES

On se propose dans cette nouvelle étape de trouver une relation transitive particulière "approximant au mieux" les m tableaux individuels C^k.
Cette relation transitive particulière est l'inconnue du problème et nous la noterons Y.

Dans tous les cas la relation binaire collective cherchée sera représentée, comme les tableaux C^k, par un tableau binaire Y, de terme général:

$y_{ij} = 1$ Si i "est en relation" avec j dans dans la relation collective
$y_{ij} = 0$ Sinon

L'expression "est en relation avec" dépend du contexte,et peut représenter l'une des 5 relations transitives définies précédemment.
Par exemple $y_{ij} = 1$ signifie i est classé avant j dans le cas où Y représente un ordre total tandis qu'il signifie i est équivalent à j si Y représente une relation d'équivalence.

Il nous faut maintenant définir ce que nous entendons par "approximer au mieux".
Nous définirons la distance d (Y , C^k) entre un tableau C^ket la relation inconnue Y.
Nous définirons ensuite ce que nous appellerons la "distance" entre l'ensemble des tableaux collectifs C^k et la relation Y.

Le problème consistera alors à rechercher la relation collective Y située à distance minimale de l'ensemble des m relations C^k.

2.3.1.Les distances associées (voir (6) p27)

a) Distance entre Y et C^k

La distance de Hölder entre la relation binaire collective cherchée Y et un tableau de relation C^k, est donnée par:

$$d_\ell^2 (Y , C^k) = \sum_i \sum_j | y_{ij} - c^k_{ij} |$$

Cette distance a les propriétés suivantes:
1) Cette distance est linéaire en terme de Y .
La linéarité est due à la séparabilité de l'expression précédente en fonction de Y .

2) Toutes les distances de Hölder sont indépendantes de la caractéristique ℓ de la métrique.
Finalement l'expression de $d_\ell^\ell(C^k, Y)$ est indépendante de ℓ et est donnée par:

$$\sum_i \sum_j c^k_{ij} - \sum_i \sum_j c'^k_{ij} y_{ij}$$

Dans le langage de la théorie des ensembles, d (C^k, Y) représente la distance de la différence symétrique entre deux ensembles.

b) Distance entre Y et tous les tableaux de relation C^k

Nous définirons la "distance" entre Y et tous les C^k par l'expression suivante:

$$\sum_{k=1}^{m} d(Y, C^k)$$

L'expression développée de cette fonction économique est donnée par:

$$\sum_i \sum_j \left(\sum_k c^k_{ij}\right) - \sum_i \sum_j \left(\sum_k c'^k_{ij}\right) y_{ij}$$

C'est à dire:

$$\sum_i \sum_j c_{ij} - \sum_i \sum_j c'_{ij} y_{ij}$$

Minimiser cette expression est équivalent à maximiser:

$$\sum_i \sum_j c'_{ij} y_{ij}$$

Cette dernière fonction économique représente le nombre total d'accords moins le nombre total de non-accords entre l'ensemble des m relations C^k et la relation Y.

En résumé, le problème consistant à trouver la relation collective Y située à "distance" minimale de l'ensemble des relations C^k est équivalent à trouver la relation collective Y , maximisant le nombre total d'accords moins le nombre total de "non-accords" avec l'ensemble des opinions émises.

2.3.2.Représentation de Y sous forme de contraintes linéaires

Nous avons vu précédemment que la relation transitive

cherchée Y était définie par un tableau binaire de terme général y_{ij}. Mais ces valeurs yij ne sont pas arbitraires et suivant la signification donnée à Y (Relations de préordre, d'ordre ou d'équivalence) les valeurs y devront vérifier un certain nombre de propriétés (transitivité, anti-symétrie, symétrie, totalité).

Ces propriétés sont définies de la façon suivante:

1) Transitivité
C'est la propriété commune à toutes les relations étudiées, elle se traduit par:

$$y_{ij} + y_{jk} - y_{ik} \leq 1 \quad \text{pour } (i \neq j \neq k)$$

2) Anti-symétrie

$$y_{ij} + y_{ji} \leq 1 \quad \text{pour } i \neq j$$

3) Symétrie

$$y_{ij} - y_{ji} = 0 \quad \text{pour } i \neq j$$

4) Totalité

$$y_{ij} + y_{ji} \geq 1 \quad \text{pour } i \neq j$$

Si Y est une relation de préordre partiel la propiété 1) est vérifiée.

Si Y est une relation de préordre total les propriétés 1) et 4) sont vérifiées.

Si Y est une relation d'ordre partiel les propriétés 1) et 3) sont vérifiées.

Si Y est une relation d'ordre total les propriétés 1), 3) et 4) sont vérifiées.

Si Y est une relation d'équivalence les propriétés 1) et 2) sont vérifiées.

Dans le cas d'une relation d'ordre total, la vérification simultanée des propriétés 3) et 4) se réduit à la relation d'égalité suivante :

$$y_{ij} + y_{ji} = 1 \quad i \neq j$$

3. LES MODELES ASSOCIES

3.1. LES MODELES BIVALENTS

Notre problème, consistant à trouver une relation collective

maximisant le nombre total d'accords moins le nombre total de non-accords, peut être représenté par un programme linéaire à variables bivalentes.
Ce programme est défini de la façon suivante:
Les variables 0-1 sont les valeurs yij de la relation Y.
La fonction économique à maximiser représente le nombre total d'accords moins le nombre total de non-accords.
Les contraintes représentent les différentes propriétés que doivent vérifier la relation Y. Il s'agit de la transitivité et selon le cas des propriétés d'antisymétrie, de symétrie ou de totalité.
La fonction économique $\Sigma\Sigma\, c'_{ij}\, y_{ij}$ est linéaire par rapport aux variables y_{ij} ainsi que les contraintes représentant les différentes propriétés de la relation Y. Il s'agit donc bien d'un programme linéaire

En fonction de la relation Y recherchée on obtient les modèles suivants:

a) Modèle pour une relation d'ordre total:

$$\text{Max} \sum_i \sum_{j\neq i} c'_{ij}\, y_{ij}$$

$$y_{ij} + y_{ji} = 1 \qquad i\neq j$$

$$y_{ij} + y_{jk} - y_{ik} \leq 1 \qquad i\neq j, j\neq k, i\neq k$$

$$y_{ij} = 0 \text{ ou } 1$$

Ce modèle a n(n-1) relations de totalité et anti-symétrie, et n(n-1)(n-2) relations de transitivité.
Ce modèle peut être réduit; division par 2 du nombre de variables et approximativement par 3 du nombre de contraintes. (Voir (13) p 139)

Voici quelques propriétés de ce modèle:

1) Ce problème est équivalent dans le cas où Y est un ordre total à la maximisation du nombre d'accords.(Voir (6) p 31)

Ce problème est équivalent également à la minimisation du nombre de désaccords ($\Sigma\Sigma\, c_{ji}\, y_{ij}$), soit enfin à la maximisation du nombre d'accords moins le nombre de désaccords ($\Sigma\Sigma$ (cij - cji)yij). (Voir (5) p 137 ,où la notion de désaccord est définie).

2) Dans le cas où les tableaux C^k ne représentent pas des ordres , même le cas d'un unique tableau de comparaisons est interessant, c'est le problème classique de la théorie des

graphes où l'on recherche un ordre à ditance minimale d'une relation binaire, d'un tournoi par exemple.

b) Modèle dans le cas d'une relation d'équivalence

$$\text{Max} \sum_i \sum_{j \neq i} c'_{ij} \, y_{ij}$$

$$y_{ij} - y_{ji} = 0 \quad i \neq j$$

$$y_{ij} + y_{jk} - y_{ik} \leq 1 \quad i \neq j,\ j \neq k,\ i \neq k$$

$$y_{ij} = 0 \text{ ou } 1$$

Ce modèle a $n(n-1)$ relations de symétrie et $n(n-1)(n-2)$ relations de transitivité. Comme dans le cas précédent le modèle peut être réduit, division par 2 du nombre des variables et division par 2 du nombre des contraintes, (voir (6) p 45).

Indiquons succinctement quelques proprités de ce modèle:

1) Contrairement au cas où Y est un ordre total, ce problème n'est pas équivalemt à la maximisation du nombre total d'accords : $\Sigma \Sigma \, c_{ij} \, y_{ij}$

ou a la minimisation du nombre total de non-accords: $\Sigma \Sigma \, (m-c_{ij}) y_{ij}$ bien

que la fonction économique $\Sigma \Sigma c'ij \, yij$ soit par définition la différence de ces deux quantités.

2) Les tableaux de similarités C^k peuvent représenter des partitions, dans ce cas particulier le problème est connu sous le nom du problème de la "partition centrale" posé par S. Régnier (voir (8)).

3) Lorsque les tableaux de similarités C^k ne représentent pas des partitions , même le cas d'un simple tableau de similarités est intéressant.C'est le problème de théorie des graphes où une relation binaire (le tableau C^k) est approximé par une relation d'équivalence (la relation Y).
Un cas particulier de ce problème est le problème de Zahn, où une relation symétrique est approximée par une relation

c) Modèles pour les autres types de relations

-la fonction économique est toujours : Max $\sum_i \sum_{j \neq i} c'_{ij} \, y_{ij}$
-les variables y_{ij} sont toujours bivalentes
-les contraintes (toutes linéaires) sont celles définies en 2-3-2.

3.2. LES PROGRAMMES LINEAIRES ASSOCIES

Si nous remplaçons les conditions $y_{ij} = 0$ ou 1, par des conditions de variation continue $0 \leq y \leq 1 \; \forall \; i \neq j$ nous obtenons, à la place des modèles bivalents précédents, des programmes linéaires continus.
Ces programmes linéaires continus peuvent être résolus par programmation linéaire , par la méthode du simplexe par exemple.

3.2.1.Programmes linéaires associés (Relation d'ordre total)

Après simplification des contraintes et élimination des variables y_{ji} pour $i \leq j$, (voir(5) p 140), on obtient le programme linéaire suivant:

$$\text{Max} \quad \sum_i \sum_{i<j} (c'_{ij} - c'_{ji}) \, y_{ij} + K$$

$$-y_{ij} - y_{jk} + y_{ik} \leq 0 \quad \text{variable duale} \quad u'_{ijk} \qquad i<j<k$$

$$y_{ij} + y_{jk} - y_{ik} \leq 1 \quad " \quad " \quad u''_{ijk}$$

$$- y_{ij} \leq 1 \quad \text{variable duale} \quad u'_{ij} \qquad i<j$$

$$y_{ij} \leq 0 \quad " \quad " \quad u''_{ij}$$

Avec

$$K = \sum_{i>j} c'_{ij}$$

Ce programme primal est sous forme "inégalité" son dual est donc sous forme "standard" et est donné par:

$$\text{Min}\ \Big[\sum_{i<j<k} u''_{ijk} + \sum_{i<j} u''_{ij} \Big] + (K)$$

$$\sum_{j<\ell} (u''_{ij\ell} - u'_{ij\ell}) + \sum_{\ell<i} (u''_{\ell ij} - u'_{\ell ij}) - \sum_{i<\ell<j} (u''_{i\ell j} - u'_{i\ell j}) + u''_{ij} - u'_{ij} = c'_{ij} - c'_{ji}$$

$$0 \le u'_{ijk} \qquad i<j<k$$

$$0 \le u''_{ijk}$$

$$0 \le u'_{ij} \qquad i<j$$

$$0 \le u''_{ij}$$

Le problème primal a $n'=n(n-1)/2$ variables et $m'=(n+1)n(n-1)/3$ contraintes. Ces deux valeurs sont inversées pour le problème dual. Les bases du primal et du dual sont identiques, leur taille est égale au nombre de variables du problème primal.

3.2.2.Programmes linéaires associés (Relation d'équivalence)

Dans ce cas on peut également simplifier les contraintes et éliminer les variables y_{ji} pour $i \le j$.
On obtient:

$$\text{Max} \sum_{i} \sum_{i<j} (c'_{ij} + c'_{ji})\, y_{ij}$$

$$-y_{ij} + y_{jk} + y_{ik} \le 1 \quad \text{variable duale } u_{ijk}$$

$$y_{ij} - y_{jk} + y_{ik} \le 1 \quad " \quad " \ u'_{ijk} \qquad i<j<k$$

$$y_{ij} + y_{jk} - y_{ik} \le 1 \quad " \quad " \ u''_{ijk}$$

$$y_{ij} \le 1 \quad \text{variable duale } u_{ij}$$

$$-y_{ij} \le 0 \quad " \quad " \ u'_{ij} \qquad i<j$$

Le problème primal est sous forme "inégalité" son dual est sous forme "standard" et est donné par:

$$\text{Min} \sum_{i<j<k} (u_{ijk} + u'_{ijk} + u''_{ijk}) + \sum_{i<j} u_{ij}$$

$$\sum_{\ell<i}(u_{\ell ij} - u'_{\ell ij} + u''_{\ell ij}) + \sum_{i<\ell<j}(u_{i\ell j} + u'_{i\ell j} - u''_{i\ell j}) + \sum_{j<\ell}(-u_{ij\ell} + u'_{ij\ell} + u''_{ij\ell}) + \ldots$$

$$\ldots + (u_{ij} - u'_{ij}) = c'_{ij} + c'_{ji} \qquad i<j$$

$$u_{ijk} \geq 0$$
$$u'_{ijk} \geq 0 \qquad i<j<k$$
$$u''_{ijk} \geq 0$$

$$u_{ij} \geq 0$$
$$u'_{ij} \geq 0 \qquad i<j$$

Le problème primal a n'=n(n-1)/2 variables et m'=n (n-1)/2 contraintes. Ces deux valeurs sont inversées pour le problème dual. Les bases du primal et du dual sont identiques; leur taille est égal au nombre n' de variables du problème primal.

3.2.3.Relation entre le modèle bivalent et le Programme linéaire associé

Nous avons détaillé ces relations dans (5) et (6).
Disons simplement que d'un point de vue théorique les deux modèles ne sont pas équivalents et que la possibilité d'obtenir une solution fractionnaire pour le programme linéaire associé existe.
Mais ce phénomène et assez rare et ne semble pas se produire dans les problèmes pratiques.

4. LA METHODE DE RESOLUTION PROPOSEE

Cette méthode est une méthode générale permettant de traiter l'ensemble des cinq problèmes différents que nous avons définis précédemment.

C'est une méthode de programmation linéaire caractérisée par les deux points suivants:

a) Résolution du modèle de programmation linéaire par le système MPSX/370
b) Utilisation du programme linéaire dual.

Ces deux caractéristiques sont extrêmement importantes puisqu'elles vont nous permettre de résoudre des problèmes réels de grande taille; c'est ce que nous allons préciser maintenant.

4.1. UTILISATION DU SYSTEME MPSX/370

(sytème de programmation mathématique IBM)

Il est nécéssaire d'utiliser un tel système pour les raisons suivantes :

1) Le système MPSX/370 possède des techniques de résolution telles que:
Décomposition LU des matrices de bases, méthode Devex pour choisir les variables entrant ou sortant de la base, traitement des cas de dégénérescence, etc..., voir (5).
Ces techniques, soit originales, soit récentes, permettent de traiter de façon particulièrement efficace de gros problèmes.

2) Ce système possède en outre un langage de contrôle étendu (ECL) en langage PL/1 (9) . Il est alors possible d'utiliser ce système dans un programme écrit en PL/1. Dans un tel programme ce système apparait sous la forme de sous programmes PL/1. Avec une telle approche la génération de l'un des cinq programmes linéaires et l'interprétation des résultats est particulièrement simple; ce qui n'est pas toujours le cas avec d'autres langages de contrôle.

4.2. UTILISATION DU PROGRAMME LINEAIRE DUAL

Cette approche conduit à des bases de taille de l'ordre n^2, au lieu de n^3 si l'on considérait le problème primal; Ce qui permet la résolution de problèmes beaucoup plus gros.

Au niveau de la résolution, les codes existant considèrent toujours le problème original sous forme standard; un problème non initialement sous forme standard est rendu standard par adjonction de variables d'écart.

Pour cette raison, il ne faut pas travailler sur le problème original sous forme inégalité. L'adjonction automatique de variables d'écart pour le rendre standard conduirait à des bases de l'ordre de n^3; ce qui enléverait tout intérêt à l'approche.
C'est le problème dual qui doit être pris comme problème de référence: il est initialement sous forme standard avec des bases de taille $n(n-1)/2$ pour les problèmes d'ordre total et d'équivalence; cette taille est multipliée par deux pour les autres cas.

La procédure "primale" du système permettra la résolution du problème dual du problème posé; la procédure "duale" celle du problème primal initial sous forme inégalité, mais avec des bases de taille $n(n-1)/2$. Ceci ne présente aucune difficulté; le générateur utilisé pouvant définir indifféremment le problème primal initial ou bien son dual.

Nous avons utilisé la procédure "primale" (résolution du problème dual) pour résoudre de nombreux problèmes concrets, environ une centaine de problèmes de classements ainsi q'une trentaine de problèmes de classification. Dans chaque cas nous avons obtenu une solution optimale bivalente.

Nous donnons ci-dessous quelques résultats sur les temps de calcul obtenus un ordinateur IBM 370/168 en utilisant le système MPSX/370. Ce temps de calcul comprend la lecture des données, la génération du programme linéaire, la résolution du problème et l'édition des résultats sous forme de rangs ou de classes:

-Pour le plus grand nombre de problèmes, ceux pour lesquels n était compris entre 10 et 30, le temps de résolution a varié de 3 à 20 secondes.

-Pour un problème d'ordre total déjà plus gros où n était égal à 36, le temps de calcul a été de 30 secondes.

-Un problème de partition centrale, de taille similaire, avec n égal 35 a été résolu en 60 secondes.

-Enfin pour un très gros problème d'ordre total, avec n égal 72, la solution optimale a été obtenue en 15 minutes environ: ce qui est un temps relativement court compte tenu de l'encombrement mémoire (2.556 contraintes, 124.392 variables et 427.705 éléments non nuls dans la matrice des contraintes). C'est la première fois qu'un problème d'une aussi grande taille a été traité par une méthode donnant la solution exacte (tout au moins à notre connaissance).

5. APPLICATIONS

Comme nous l'avons dit au paragraphe précédent , nous avons traité de nombreux problèmes réels par les méthodes et algorithmes que nous avons présentés; environ une centaine de problèmes de classement et une trentaine de problèmes de classification.

Certains des problèmes que nous avons traités sont ou seront prochainement présentés dans des publications. (Voir (1), (5))

A titre d'exemple, nous allons appliquer la méthode générale sur un problème de classification, extrait de la revue METRA N°3 Vol.13 (1974). Il s'agit d'un article de J.M.Tardieu et J.P.Vicario sur "l'application des techniques multivariées à l'image de marque en marketing industriel".

Nous allons traiter le problème de classification de 10 machines à

écrire, présenté dans cet article.

On a demandé à 57 personnes d'indiquer si les machines qu'on leur présente sont "semblables" ou non. Les réponses obtenues (par comparaisons) sont reportées dans le tableau de similarités suivant:

TABLEAU DES SIMILARITES ENTRE MACHINES A ECRIRE

	IBM	ADLER	FACIT	OLIVE.	HERMES	M.TYPE	REMIN.	JAPY	OLYMP	TRIUM.
IBM		15	11	32	1	34	4	6	17	7
ADLER	15		27	35	27	27	20	24	30	15
FACIT	11	27		25	25	20	34	25	17	15
OLIV.	32	35	25		22	44	13	22	30	11
HERM.	1	27	25	22		21	28	43	20	13
M.TYPE	34	27	20	44	21		18	27	21	8
REMIN.	4	20	34	13	28	18		31	28	13
JAPY	6	24	25	22	43	27	31		30	15
OLYMP.	17	30	17	30	20	21	28	30		12
TRIUM.	7	15	15	11	13	8	13	15	12	

A l'intersection de la ligne i et de la colonne j on trouve le nombre de personnes jugeant i et j semblables.

La résolution du problème $\text{Max} \sum_i \sum_{j \neq i} c'_{ij} y_{ij}$ sous les contraintes définies en 3-2-2 aboutit au résultat suivant:

Valeur de la fonction économique à l'optimum:

184

Nombre de classes:

5

Classe 1 : IBM, OLIVETTI, Mach.Type

Classe 2 : ADLER, OLYMPIA

Classe 3 : FACIT, REMINGTON

Classe 4 : HERMES, JAPY

Classe 5 : TRIUMPH

6. REFERENCES

1 AYACHE G.,MARCOTORCHINO J.F., MICHAUD P. "Démocratie Française et la presse, Techniques d'Analyse de la Réception d'un message", Revue Française de Communication, No 1, 30, (1978).

2 BARTHELEMY J.P., MONJARDET B., "Ajustement et Résumé de Données Relationnelles: Les Relations Centrales", présenté à ce même Symposium.

3 DIDAY E., "Nouvelles Méthodes et Nouveaux Concepts en Classification Automatique et Reconnaissance des Formes, Thèse de Doctorat d'Etat, Université de Paris VI,(1972).

4 JACQUET-LAGREZE E., "l'Agrégation des Opinions Individuelles", Informatique et Sciences Humaines, No 4, (1969).

5 MARCOTORCHINO J.F., MICHAUD P., "Optimisation en Analyse Ordinale des Données", Masson, Paris (1979).

6 MICHAUD P., MARCOTORCHINO J.F., "Optimization in Ordinal Data Analysis", (article présenté pour une partie au Symposium International d'Optimisation discrète, Vancouver (1977),) et disponible dans Rapport Technique No 1, IBM France Scientific Centre, (1978).

7 MONJARDET B., "Tournois et Ordres Médians pour une Opinion", Mathématiques et Sciences Humaines, No 43, 55, (1973).

8 REGNIER S., "Sur quelques Aspects Mathématiques des Problèmes de Classification Automatique", I.C.C., Bulletin No 4, 175, (1965).

9 MPSX/370, "IBM Mathematical Programming System Extended/370", General Information Manual, GH 19-1090.

T3

OPTIMIZATION IN EXPLORATORY DATA ANALYSIS

(Preference and similarity aggregation)

F.MARCOTORCHINO P.MICHAUD

I.- INTRODUCTION

The aim of this communication is to present a new approach for solving a certain number of "exploratory" data analysis problems, by optimization.
We will present here two interesting applications of this general approach: first, the "Preference aggregation problem" and second, the "Similarity aggregation problem".

-The "preference aggregation problem", also known as the "collective choice problem", consists of finding a collective ranking of "n" objects, summarizing in "the best way"individual preferences given by m "judges" or "criteria".
Since the origins (the Condorcet's work on voting procedures in 1785) and more recently, according to the restricting conditions of the Arrow's theorem, this problem was still considered as a very difficult one.
The method proposed in this paper gives a solution, which is in a certain way the "best" among all the possible solutions.

-The "similarity aggregation problem" can be defined as the research of an "optimal collective partition" (related to cluster analysis), summarizing or approximating in the "best way" individual similarities upon "n" objects, given by m "judges" or obtained from m "criteria".
This problem is less known than the previous one, but in the rare articles on this subject in the scientific literature, it is considered as very difficult, even impossible to solve exactly, (see <4>,<16> and <17>).

In that case also, we will give a method, similar to the previous one, allowing to obtain an exact solution in a reasonnable amount of time.

There exist only "branch and bound" methods for solving the first problem; but these methods, because of the too large amount of time required, do not permit to solve problems of size greater than "n"=.20.
For the "similarity aggregation problem", there didn't exist any exact method until today; only heuristics are used, like the "dynamic clustering method" of E. Diday and other related algorithms.

These two problems, a priori very different, have the following caracteristics in common:
a-Both problems can be written under a linear programming form.

b-Each problem deals with the search of a particular binary relation (a linear order "ranking" in the first case, an equivalence relation "partition" in the second case) approximating "m" individual binary relation in the "best way". (The meaning of the terms "best way" will be explained later).

c-These two problems are strongly related to the "Condorcet's majority rule".

d-The data are transformed into a relational presentation.

II.- DATA REPRESENTATION

Each initial binary relation, defining the individual data, given by the judge or criterion number k, will be represented by a "comparison" or "relational" tableau C^k. This comparison tableau is a (zero-one) tableau defined as follows:

$c^k_{ij} = 1$ If i "is in relation with" j for the "judge" k; i.e if iRj

$c^k_{ij} = 0$ Otherwise.

Conventionally the values c^k_{ij} with i=j will not be considered here.

According to the type of problem and the nature of the data, the meaning of the cij values will be different.

For example:

-i is "in relation with" j means: i is "preferred to" j if C^k is a preference tableau.

-i is $_k$"in relation with" j means: i is "similar to" j if C^k is a similarity tableau.

For each binary relation represented by a C^k tableau, one can associate another comparison tableau, noted C'^k ; this new tableau is a (-1,+1) tableau defined by:

$c'^k_{ij} = 1$ If i is "in relation with" j for the judge k

$c'^k_{ij} = -1$ Otherwise

The values c'^k_{ij} can be expressed in terms of the c^k_{ij} by:

$$c'^k_{ij} = 2\, c^k_{ij} - 1$$

The $C'^k$and the C^ktableaux may be obtained directly or indirectly by means of annex tableaux (ranking, rating or distances tableaux etc.).

Remark: Any binary relation can be represented by the C'^k and C^k tableaux; in particular complete orders or complete quasi-orders, partial orders or partial quasi-orders, semi-orders, tournaments, equivalence relation etc..(see <12> for further comments).

Summing up all the individual tableaux C^k and C'^k, representing the given data, one obtains collective tableaux C and C' by:

$$C = \sum_k C^k \quad \text{and} \quad C' = \sum_k C'^k \quad \text{and} \quad C' = 2C - m$$

the (i,j)th element of C is:

$$c_{ij} = \sum_k c^k_{ij}$$

The (i,j)th element of C' is:

$$c'_{ij} = \sum_k c'^k_{ij}$$

Remark: The meaning of the C' tableau is of interest because of its connection with the "Condorcet's majority rule". Indeed,

c'_{ij} = The number of judges putting i in "relation with" j minus the number of judges doing not.

Hence,
$c'_{ij} > 0$ implies : a majority of judges consider that i is in relation with j.
This is an extension to general comparison tableaux of the Condorcet's majority rule, available for preference or voting comparison tableaux C'.

III.-DEFINITION OF THE GENERAL AGGREGATION PROBLEM

3.1- Definition of the unknown collective relation

For both problems (preference and similarity aggregation), we are looking for a particular transitive and collective relation either a complete order or an equivalence relation (partition) respectively, "approximating in the best way" the m individual binary relations C^k. This transitive relation we are looking for, will be denoted Y in the rest of this paper.
In all cases this unknown collective binary relation Y will be represented, as the C^k tableaux, by a binary tableau Y built as follows:

$y_{ij} = 1$ if i is "in relation with " j in the collective result

$y_{ij} = 0$ Otherwise.

The terms "is in relation with " depend upon the context and may have different meanings according to the nature of Y.

For example:
$y_{ij} = 1$ means: i is ranked higher than j if Y represents a complete order.
$y_{ij} = 1$ means: i is in the same cluster as j if Y is an equivalence relation.

But in the general case Y may represent any other kind of binary relation : quasi-order, semi-order etc.. (see <9> and <12>).

3.2- Definition of an objective function

We now have to define the meaning of the terms:"approximating in the best way".
For that, let us consider the following function of Y:

$$F(Y) = \sum_{ij} \sum c'_{ij} \, y_{ij}$$

where c'_{ij}, as we have already seen, represents the number of judges putting i in relation with j minus the number of judges doing not; or with another interpretation c'_{ij} = the number of judges in agreement with iRj minus the number of judges not in agreement with iRj.

This function F(Y) is trivially maximum for Y defined by:

$y_{ij} = 1$ if $c'_{ij} \geq 0$

$y_{ij} = 0$ otherwise.

We have already seen that $c'_{ij} > 0$ implies "a majority of judges consider i in relation with j". Consequently Y is nothing else that the result of the direct application of the "Condorcet's majority rule" to the C' comparison tableau.
It will be noted here that Y is given without constraints on it, and may consequently represent any binary relation. However except for peculiar structures of the C' tableaux (see for example the Black's conditions in <1>, and other conditions in <12>), this solution will not give , a priori, respectively either a complete order or an equivalence relation. We shall precise these points later.

In order to explain the relationship between the expression "approximating in the best way" and the "Condorcet's majority rule", we have to point out how the "objective function" or "criterion" F(Y) derives from metric properties related to an approximation point of view.

a) First we must define the distance $d(Y,C^k)$ between an individual tableau C^k and the unknown relation Y.

$$d(Y,C^k) = \sum_i \sum_j | y_{ij} - c^k_{ij} |$$

This expression leads to the following linear form (see <9 > p27)

$$d(Y,C^k) = \sum_i \sum_j c^k_{ij} - \sum_i \sum_j c'^k_{ij} y_{ij}$$

Remark: This distance between C^k and Y is in fact the "symmetric difference distance" between the relation C^k and Y.

b) In order to take into account the idea of approximation, the "distance" between Y and all the C^k tableaux will be obviously defined by:

$$F'(Y) = \sum_{k=1}^{m} d(Y,C^k)$$

it leads also to a linear form in y_{ij}, depending only upon the c_{ij} and c'_{ij} values. (See <9 > p28)

$$F'(Y) = \sum_i \sum_j c_{ij} - \sum \sum c'_{ij} y_{ij}$$

Saying that "Y approximates in the best way the Ck tableaux " is equivalent to say that Y minimizes F'(Y).
But we can see that:

$$F'(Y) = Cte - F(Y)$$

hence we can deduce these two prominent results:

1) The objective function is linear in terms of Y.

2) The problems:

$$\min_{Y} F'(Y) \quad \text{(metric presentation)}$$

$$\max_{Y} F(Y) \quad \text{(Condorcet's rule presentation)}$$

are equivalent and if we don't suppose any constraints on Y are obviously solved by putting:

$$y_{ij} = 1 \quad \text{if} \quad c'_{ij} \geq 0$$

$$y_{ij} = 0 \quad \text{otherwise.}$$

But unfortunately, the unknown collective relation Y must verify a certain number of constraints due to the fact that we want either a complete order or an equivalence relation.

3.3- Representation of Y by linear constraints

If the unknown binary relation Y we are looking for is a complete order, Y must be: (transitive, asymmetric, complete)

If Y is an equivalence relation, Y must be (transitive, symmetric).

Then, we have again an interesting result:

All these properties can be formulated in terms of linear expressions of the y values . (See <6>,<9> and <12> for more details)

1)Transitivity

$$y_{ij} + y_{jk} - y_{ik} < 1 \quad \forall\ (i,j,k) \text{ all different}$$

This inequality means that if iRj and jRk then iRk for the relation Y.

2)Symmetry

$$y_{ij} - y_{ji} = 0 \quad \forall\ (i \neq j)$$

it means that if (for a clustering problem) i is in the same cluster as j, then j is in the same cluster as i.

3)Asymmetry

$$y_{ij} + y_{ji} \leq 1 \quad \forall\ (i \neq j)$$

This inequality means that at the most iRj or jRi is true for the relation Y.

4) Completness

$$y_{ij} + y_{ji} \geq 1 \quad \forall\ (i \neq j)$$

This inequality means that at least iRj or jRi is true for the relation Y.

If Y is a complete order, Y must verify simultaneously 3) and 4) this leads to the new relation number 5)

5) $y_{ij} + y_{ji} = 1 \quad \forall\ (i \neq j)$

This relation means that (in a complete order) either i is ranked higher than j or j is ranked higher than i.

We now can deduce a general interpretation of the preference and similarity aggregation problems :
These two problems can be considered as the verification of the "Condorcet's majority rule " under the constraints 1) and 5) for the first problem and 1) and 2) for the second.

IV.-THE ASSOCIATED LINEAR MODELS

4.1- The bivalent models

It derives from the previous results that the preference and similarity aggregation problems can be formulated as linear integer programming problems in (0-1) variables.

The (0-1) variables are the values y_{ij} of the collective relation Y we are looking for.

The objective function to maximize, already defined, is F(Y) and it is a linear function of Y.

The constraints represent the various properties that the relation Y must verify (transitivity, plus, according to the case, either asymmetry and completness or symmetry). We have already seen that these constraints are linear expressions of Y.

According to the type of Y, we obtain the following linear models:

a) Model for a complete order (ranking)

$$\text{Max} \sum_i \sum_j c'_{ij}\, y_{ij}$$

$$y_{ij} + y_{jk} - y_{ik} \leq 1 \quad \forall\ i,j,k \text{ all different}$$

$$y_{ij} + y_{ji} = 1 \quad \forall\, i \neq j$$

$$y_{ij} = 0 \text{ or } 1$$

b) Model for an equivalence relation (partition)

$$\text{Max} \sum_i \sum_j c'_{ij}\, y_{ij}$$

$$y_{ij} + y_{jk} - y_{ik} \leq 1 \quad \forall\, i,j,k \text{ all different}$$

$$y_{ij} - y_{ji} = 0 \quad \forall\, i \neq j$$

$$y_{ij} = 0 \text{ or } 1$$

(One can find in <9>, and especially in <12> a certain number of linear models associated to more complicated Y relations).

The number of constraints of these problems is of order n-cubed, while the number of variables is of order n-squared.

Remark: One can reduce the size of these models, (see <6> p.139 , and <9> p.45), but the order of magnitude remains identical.

c) Some properties of the previous models

1) The case of a complete order relation

i) The maximization of the objective function F(Y) is equivalent in that case to the maximization of the number of "agreements" defined by:

$$\sum_i \sum_j c_{ij}\, y_{ij}$$

it is also equivalent to the minimization of the number of "desagreements" defined by:

$$\sum_i \sum_j c_{ji}\, y_{ij}$$

or it is equivalent to the maximization of the number of agreements minus the number of desagreements:

$$\sum_i \sum_j (c_{ij} - c_{ji})\, y_{ij}$$

(See <6> and <9> for more details)

ii) When the C^k tableaux do not represent complete rankings, even the case of an unique C^k tableau is interesting, it corresponds to a classical problem of graph theory consisting of finding a complete order located at a minimal distance of a given binary relation, for example a tournament.

iii)The particular case where the C^k tableaux represent rankings is known as the "Kemeny's problem", but no solving method as been proposed by this author (see <3>).

2)The case of an equivalence relation

i)In spite of the resemblance between the two problems, the maximization of the total number of agreements

$$\sum_i \sum_j c_{ij} y_{ij}$$

is not equivalent to the maximization of F(Y), in that case. This leads to the trivial solution $y_{ij} = 1 \quad \forall \ (i \neq j)$, i.e all the objects are in the same cluster.

ii)If the individual similarity tableaux C^k represent partitions, the similiraty aggregation problem is known as the "Regnier's problem" (see <16>).The Regnier's problem is then a particular case of the similarity aggregation problem.

iii)When the individual C^k tableaux do not represent partitions, even the case of a single similarity tableau is of interest. It is a classical problem of graph theory where a binary relation (the C^k tableau) is approximated by an equivalence relation (the Y tableau). A particular case of this problem is known as the "Zahn's problem", (a symmetric relation must be approximated by an equivalence relation). (See <18>).

4.2- The associated continuous programs

If we replace the condition $y_{ij} = 0$ or 1 by $0 \leq y_{ij} \leq 1$, one obtains a continuous linear program. Unfortunately the number of constraints being of order n-cubed, it is impossible to solve this linear program even for small values of n.
But symmetricaly the dual of this linear program has only an order of n-squared constraints (with n-cubed variables)and can be consequently solved by linear programming in a reasonnable amount of time.(See <6>,<9>,<11>).

Remark: For the linear continuous problem the optimal

solutions of the "primal" problem and of the "dual" problem are identical. But in general the bivalent linear program and the primal continuous linear program are not necessarily equivalent. The possibility of finding fractionnal solutions exists for the primal continuous problem as pointed out by A.J.Hoffman, E.L.Johnson and ourselves in <9>. In order to avoid this problem, we have the opportunity to use cutting planes methods in addition to the linear continuous program. However, and it is a practical result very satisfactory, for all the practical examples we have solved (more than 130), the optimal solution of the continuous linear program (obtained without using cutting planes) was always in (0-1) and was also the optimal solution of the bivalent linear program. (For more details about these points see <6> and <9>).

4.3- Presentation of the results

1) The case of a complete order

If Y is an optimal solution of the first linear program, we deduce the optimal ranking r_i of the object i as follows:

$$r_i = \sum_j y_{ji} + 1 = n - \sum_j y_{ij}$$

2) The case of an equivalence relation

If Y is an optimal solution of the second linear program, we deduce trivially the optimal partition by:

$y_{ij} = 1$ i is in the same cluster as j;
the number of this cluster is the one corresponding to the smallest index i.

4.4- The proposed method

The proposed method is a linear programming method caracterised by the following two points:

a) Resolution of the models by the IBM MPSX/370 system

b) Use of the linear programs under inequality form.

The reasons of this choice are given in <6> and <9>.

We now give some computing times for exact solving of the dual continuous linear program by using the IBM MPSX/370 system with the extended control language ECL on a IBM 370/168 (these times include generation, solving and edition of the results under the appropriate form):

EXAMPLES	SIZE(n)	TYPE	TIMES(in sec.)
1) Archaeology	21	P	7
2) Medicine	23	P	13
3) Linguistics	28	R	11
4) Medicine	29	P	22
5) Movies ranking	36	R	22
6) Psychology	35	P	36
7) Sociology	40	R	31
8) Linguistics	45	R	58
9) Esthetics	54	R	117
10) Politics	54	P	220
11) Sociology	72	R	157
12) Linguistics	77	R	330

(The capital letters P and R indicate that the result is either a "Partition" or a "Ranking").

Although the computing time is not negligible in the last example, do not forget that the number of constraints was 2926 and the number of variables 146300 for this problem.

V.-EXAMPLE OF APPLICATION

In order to illustrate how the method works, we now give a small example, adapted from Hole and Shaw, we have selected in an article of J.C.LINGOES (see <5>). The entries of the original raw Data matrix are the frequencies of 21 types of artifacts in 11 archaeological sites.

The following names of these oriental sites are ordered from oldest to most recent:

SABZ, KHAZIN3, KHAZIN2, KHAZIN|, MEHMEH3,MEHMEH2, MEHMEH1, MUSIYAN, BAYAT3, BAYAT2, BAYAT1.

The result we are looking for in that example is a "partition" of the eleven sites. The latent purpose is to see if the "serial" aspect is preserved in the expected result.

We directly give here the (11X11) similarity C tableau, associated to the original (11X21) data matrix:

	SABZ	KHA3	KHA2	KHA1	MEH3	MEH2	MEH1	MUSI	BAY3	BAY2	BAY1
SABZ		9	10	7	7	6	8	5	8	5	5
KHA3	9		9	7	4	7	3	10	6	9	9
KHA2	10	9		12	13	9	9	11	9	5	8
KHA1	7	7	12		13	7	8	10	8	7	6
MEH3	7	4	13	13		10	11	7	8	4	5
MEH2	6	7	9	7	10		13	11	12	10	9
MEH1	8	3	9	8	11	13		11	16	10	8
MUSI	5	10	11	10	7	11	11		15	12	12
BAY3	8	6	9	8	8	12	16	15		13	9
BAY2	5	9	5	7	4	10	10	12	13		14
BAY1	5	9	8	6	5	9	8	12	9	14	

= C

We have already seen that we can associate another comparison tableau to the C tableau: the C' tableau. This C' tableau is equal to (2C-21), since there are 21 types of artifacts (i.e 21 "judges" in our terminology).
With the same ordering of the rows and columns as for C, the C' tableau is given by:

	-3	-1	-7	-7	-9	-5	-11	-5	-11	-11
-3		-3	-7	-13	-7	-15	-1	-9	-3	-3
-1	-3		3	5	-3	-3	1	-3	-11	-5
-7	-7	3		5	-7	-5	-1	-5	-7	-9
-7	-13	5	5		-1	1	-7	-5	-13	-11
-9	-7	-3	-7	-1		5	1	3	-1	-3
-5	-15	-3	-5	1	5		1	11	-1	-5
-11	-1	-1	-1	-7	1	1		9	3	3
-5	-9	-3	-5	-5	-3	11	9		5	-3
-11	-3	-11	-7	-13	-1	-1	-3	-5		7
-11	-3	-5	-9	-11	-3	-5	-3	-3	7	

= C'

The optimal value of F(Y) is:

F(Yopt)= 100

The optimal solution Yopt is :

Yopt =

1	0	0	0	0	0	0	0	0	0	0
0	1	0	0	0	0	0	0	0	0	0
0	0	1	1	1	0	0	0	0	0	0
0	0	1	1	1	0	0	0	0	0	0
0	0	1	1	1	0	0	0	0	0	0
0	0	0	0	0	1	1	1	1	0	0
0	0	0	0	0	1	1	1	1	0	0
0	0	0	0	0	1	1	1	1	0	0
0	0	0	0	0	1	1	1	1	0	0
0	0	0	0	0	0	0	0	0	1	1
0	0	0	0	0	0	0	0	0	1	1

The optimal partition is then given by:

Cluster number 1 : SABZ

Cluster number 2 : KHAZIN3

Cluster number 3 : KHAZIN2, KHAZIN1, MEHMEH3

Cluster number 4 : MEHMEH2, MEHMEH1, MUSIYAN, BAYAT3

Cluster number 5 : BAYAT2, BAYAT1

The optimal solution Yopt we have found is very satisfactory. This can be stated by the following points:

1)According to the "classification criterion" (the "Condorcet's majority rule under constraints"), the more small the number of violations of the majority rule is, the more significant the solution is.

A violation of the majority rule can occur if:

either: $c'_{ij} < 0$ and $y_{ij} = 1$ Type (1)

or: $c'_{ij} > 0$ and $y_{ij} = 0$ Type (2)

Since in our case, the total number of violations of the majority rule is equal to 5 (for the upper triangular part of the C' tableau), (see the encircled cases), the optimal partition found is very significant.

Similarly, we can define two types of goodness of fit:

either: $c'_{ij} \geq 0$ and $y_{ij} = 1$ Type (1)

or: $c'_{ij} < 0$ and $y_{ij} = 0$ Type (2)

2) The intra-homogeneity of the classes is very good since the number of type 1 majority's violations is equal to 0.

3) Let us define the following quantities for any solution Y:

N1=Absolute value of type 1 errors $= -(\sum\sum_{ij} c'_{ij}\, y_{ij}) \mid c'_{ij}<0$

P0=Value of type 2 errors $= \sum\sum_{ij} c'_{ij}\, (1 - y_{ij}) \mid c'_{ij}>0$

P1=Value of type 1 goodness of fit $= \sum\sum_{ij} c'_{ij}\, y_{ij} \mid c'_{ij}>0$

N0=Absolute value of type 2 goodness of fit $= -(\sum\sum_{ij} c'_{ij}(1 - y_{ij}))$ with $c'_{ij}<0$

From the preceding expressions we can deduce the following "coherence index" valid for any "partition" Y:

$$Co(Y) = 1 - ((N1+P0)/ (N1+P0+P1+N0))$$

In other words this index is equal to:

Co(Y) =(Value of goodness of fit)/$\sum_i \sum_j |c'_{ij} + c'_{ji}|$

This index can vary from 0 to 1

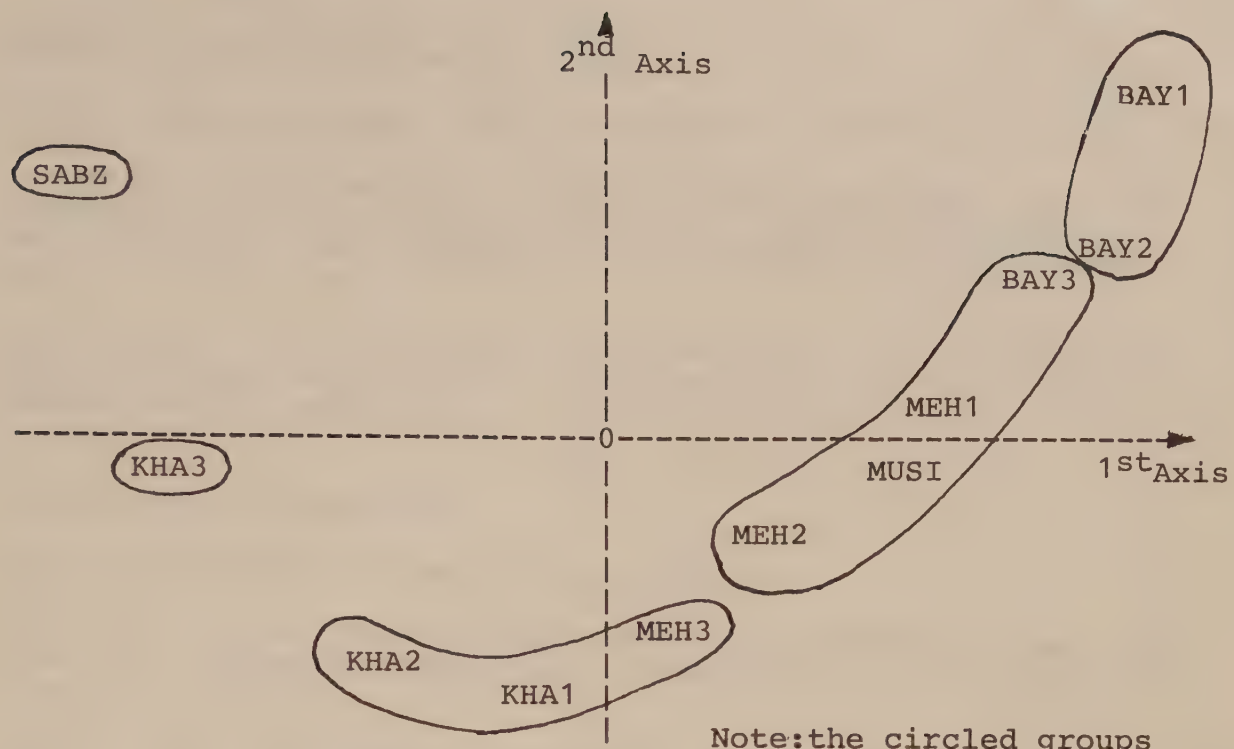
Remark: This coherence index is only available for symmetric relations and is a particular case of an obviousand general coherence index, defined for any type of relation. (See the general definition of Co(Y), given by P.Michaud, in <14>).

In our case, Co(Yopt) = 1 - 26/606= 0.9571

It is a very good "coherence", and this completes the proof that the optimal partition is formally significant.

4) On the other hand, for the concrete interpretation it is interesting to see here, that the latent archaeological "seriation" of the sites is almost preserved in the optimal partition (except the separation between BAY3 and the other Bayat sites).

And this result is also closely in agreement with the following U-shaped diagram, obtained by "correspondances factor analysis" of the original raw data matrix:



Note:the circled groups correspond to the clusters of the Optimal Partition.

VI.-CONCLUDING REMARKS

In order to conclude, we now give a list of key points related to the general approach we have proposed.

1) Common properties

a) Both problems (preference and similarity aggregation) are relational problems.

b) These two problems can be written under a linear programming form, since:

The objective function is linear in terms of the unknown Y
The constraints are linear in terms of Y

c) In both cases the problem can be interpretated as the "Condorcet's majority rule" under constraints.

d) The relational presentation of the data allows us to take into account some paradoxal data like preferences affected by a "Condorcet effect" or similarities affected by a "chain paradox", in addition to the other classical types of raw data.

e) A certain number of problems cited in the scientific literature are particular cases of the general problem we have defined. For example: the collective choice problem, Kemeny's problem, Remage-Thompson's problem, Regnier's

problem, Zahn's problem etc...

f) For both problems, we have defined indexes allowing to check:
-a priori,the quality of the data associated to the potential notions of "rankability" or "partitionability".
-a posteriori, the validity of the optimal solutions found. (see <9> , <8> and <14>).

g)For both problems, we have defined very efficient heuristic methods, giving the optimal solution in most of cases.(See <6> p145..).

2)Respective properties

a) The preference aggregation problem.

When the solution we are looking for is a complete order, we have the following properties in addition to the previous list:

i)For the preference aggregation problem, the optimal solutions are independent of the rating scales used by the "judges", only the associated ordering is taken into account.

ii)If no violation of the majority rule occurs, it is known that the solution obtained by the Condorcet's rule is optimal. And therefore this solution verifies all the Arrow's conditions.In the opposite case (if there are some violations), the solution obtained by the proposed method is the best possible regarding the Arrow's conditions.

iii)The heuristic we have developped for the preference aggregation problem has very interesting mathematical properties related to the "permutations polyhedron".(See<6>).

b) The similarity aggregation problem

When we are looking for an optimal partition we have the following extra properties:

i)In that case we have note to choose, a priori, the number of clusters of the expected partition, on the contrary of the other non-hierarchical clustering methods (Hall and Ball, Diday etc..).

ii)If the data are (zero-one) entries (presence-absence), the specialists often use "similarity indexes" like the Sokal and Sneath's index, the Rogers and Tanimoto's index, the Sokal and Michener's index etc...It is very interesting

to state that the relational property of the problem and its close relationship with the Condorcet's majority rule, allow us to consider that all these indexes are only simple variations of the majority rule (as pointed out by F.Marcotorchino in <8>).

iii)According to the constraints generally placed on classification procedures, as mentionned by D.M.Jackson in <2>:

We can assert that:
-The obtained classification (the optimal partition) is "stable". That is, this classification is not grossly affected by small changes in the data.
-This obtained partition is independent of the "labeling" of the n objects to classify.
-This proposed classification procedure is independent of"scale". That is the solution is unaffected by multiplication of the similarity matrix C' by a positive constant.

Of the four conditions proposed by Jackson, our method does not verify only the first: the idea of "well defined"(a classification procedure must supply a "single" result). But to our point of view this condition is not as relevant as the others.

VII.- REFERENCES

<1> D.BLACK,The Theory of Comittees and Elections, Cambridge University Press, Cambridge (1958).

<2> D.M.JACKSON, Classification, Relevance, and Information Retrieval,in Advances in Computers Edited by Franz L.Alt and al. Vol 11, Academic Press, New York, (1971).

<3> J.KEMENY and L.SNELL, Mathematical Models in the Social Sciences, The M.I.T. Press, Cambridge USA,(1962)

<4> I.C.LERMAN, Les Bases de La Classification Automatique, Gauthier Villars, Paris, (1970).

<5> J.C.LINGOES, A general Nonparametric Model For Representing Objects and Attributes in a Joint Metric Space, in Archeologie et Calculateurs International Symposium, Marseille Avril 1969, Editions du CNRS, Paris (1970).

<6> F.MARCOTORCHINO and P.MICHAUD, Optimisation en Analyse Ordinale des Donnees, Masson, Paris, (1979).

<7> F.MARCOTORCHINO and P.MICHAUD, Methodes d'Analyse Ordinale des Donnees, in "proceedings" of the CERME Symposium, Nice September 1978, Publication de l'IAE de Nice,(1978)

<8> F.MARCOTORCHINO and P.MICHAUD, Optimisation en Classification Automatique, (to Appear in Revue de Statistique Appliquee, Dunod , Paris (1980)).

<9> P.MICHAUD and F.MARCOTORCHINO, Optimization in Ordinal Data Analysis, IBM-France Scientific Center Technical Report 001, (1978).(An important Part of this Paper has been presented at the International Symposium on Discrete Optimization, Vancouver August 1977).

<10>P.MICHAUD and F.MARCOTORCHINO, Optimisation en Analyse et Agregation d'Opinions, IBM-France Scientific Center Technical Report 002, (1979).(Papers presented at The International Symposium on Preference Aggregation, Aix en Provence Juin 1979).

<11>P.MICHAUD and F.MARCOTORCHINO, Optimisation en Analyse des Donnees Relationnelles, in the "Proceedings" of the International Symposium on Data Analysis, sponsored by IRIA in Versailles, October 1979. (To Appear by North Holland 1980).

<12>P.MICHAUD and F.MARCOTORCHINO, Modeles d'Optimisation en Analyse des Donnees Relationnelles, in Mathematiques et Sciences Humaines N 67, pp7-38, Dunod Paris, (1979).

<13>P.MICHAUD and F.MARCOTORCHINO, Optimization in Data Analysis, in the "proceedings" of the 10th International Symposium on Mathematical Programming, Montreal August 1979, (to appear 1980).

<14>P.MICHAUD and F.MARCOTORCHINO, Une nouvelle Methode d'Agregation d'Opinions, to appear in Les Annales de l'Institut National de la Statistique, Dunod Paris (1980).

<15>MPSX/370, IBM Mathematical Programming System Extended 370, General Information Manual, GH19-1090.

<16>S.REGNIER, Sur Quelques Aspects Mathematiques des Problemes de Classification Automatique, I.C.C Bulletin, Rome, (1965).

<17>S.REGNIER, Stabilite d'un Operateur de Classification,

Math. et Sciences Humaines, N 60, pp21-30, Dunod Paris, (1977).

<18>C.T.ZAHN, Approximating Symmetric Relations by Equivalence Relations, SIAM J. of Applied Math., N 12, p840, New York,(1964).

Compagnie IBM France

Réalisé sur les presses de l'Imprimerie d'Entreprise
Usine de BOIGNY – FRANCE

1–4142–05–81